DIAL
Digital access to libraries

"Understanding social dynamics through big data"

Deville, Pierre

Abstract

Data are everywhere. They pervade our world. From e-mails we send, online status we post or friends we call, to credit cards we swipe or papers we cite, most of our everyday actions leave digital traces. As our ability and capacity to measure natural and social phenomena is rapidly increasing at an unprecedented scale, we witness an exponential growth of all these digital traces. This growing digital information is what we call Big Data; Data that we generate and acquire far more rapidly than the rate at which we process, analyse and exploit it. In science, the ability to collect and analyse massive amounts of data traces have fuelled numerous advances and unambiguously transformed many research fields. But nowhere are these advances more important than in the study of social systems. Indeed, the flood of data capturing activities of individuals enables an entirely new scientific approach for social analysis, which this thesis aims at illustrating. More particularly, our contribution...

Document type : *Thèse (Dissertation)*

Référence bibliographique

Deville, Pierre. *Understanding social dynamics through big data.* Prom. : Blondel, Vincent

Available at:
http://hdl.handle.net/2078.1/165615

[Downloaded 2019/04/19 at 03:16:35]

Université catholique de Louvain
Ecole Polytechnique de Louvain
Institut ICTEAM
Pôle en ingénierie mathématique

Understanding Social Dynamics through Big Data

Pierre Deville

Thesis submitted in partial fulfillment
of the requirements for the degree of
Docteur en Sciences de l'ingénieur

Advisor: Prof. Vincent Blondel

Dissertation committee:

Prof. Vincent Blondel (Université catholique de Louvain)
Prof. Raphaël Jungers (Université catholique de Louvain)
Prof. Renaud Lambiotte (Université de Namur)
Asst. Prof. Roberta Sinatra (Northeastern University, USA)
Dr. Zbigniew Smoreda (Orange Labs, France)
Prof. Paul Van Dooren (Université catholique de Louvain)

Acknowledgements

A few years ago would you have told me I would become a scientist, I would have probably laughed. Not that I thought science was a joke, but because of the total misconception I had about it. Fortunately, over the last four years, I had the chance to collaborate with truly amazing people from a broad variety of places and disciplines, who helped me perceive what science really is. Only fools never change their minds as they say. I would like to thank all these people here.

First and foremost, I would like to thank my advisor Vincent Blondel who gave me the tremendous opportunity to go to MIT when I was still a student. Without a doubt, this has deeply transformed and broadened my vision of what I though was research at that time. This also led me to start a thesis under his supervision where I had the chance to benefit from a total independence regarding my collaborations or projects and for which I am more than grateful.

This thesis has been reviewed by my dissertation committee which I would like to thank for the many insightful comments they gave me during my Ph.D. In particular, I would like to thank Roberta Sinatra who has been one of my main mentors when I was in Boston and who guided me throughout all these years. I would also like to thank Renaud Lambiotte and Paul Van Dooren for being in my supervisory committee and for the friendly interactions we had during these past four years. I am also thankful to Raphaël Jungers for guiding me through the final process of this thesis and for chairing this committee. Finally, I would like to thank Zbigniew Smoreda for the many insightful discussions we had since I was a student and for helping me handling some of the data used in this work.

Most of my research has been done in collaboration with the Center for Complex Network Research in Boston where I had the chance to stay during most of my Ph.D. I would like to sincerely thank László Barabási for inviting and supervising me during all these years in his lab as well as for his insightful advice on my research. Of course, this

research stay would not have been the same without all the amazing people I met inside and outside of the lab. In particular, I would like to thank Dashun Wang for supervising my work during my first years in the lab as well as Joe, Brett, Sarah and Suzanne for their continuous IT and administrative help.

I would also like to thank all my colleagues from the Euler building for the great atmosphere and the many activities they organised when I was in Belgium. In particular, I would like to thank the administrative staff for their precious help as well as Etienne Huens who helped me deal with the many datasets I had to work with.

During my Ph.D., I was financially supported by the Fond de la Recherche Scientifique - F.N.R.S., which I would like to thank. I have also been supported by the Action de Recherche Concerté (ARC) Large Graphs and Networks as well as the IAP DYSCO (Interuniversity Attraction Poles programme for Dynamical systems, control and optimization) managed by the Belgian Science Policy Office (BELSPO).

Finally, I would like to deeply thank my family, and more particularly my wife Angélique, for all their continuous support and encouragement.

Contents

Chapter 1

General Introduction

Data are everywhere. They pervade our world. From our everyday activities such as the e-mails we send, friends we call, credit cards we swipe or papers we cite, to experiments we make such as astronomical measurements, genomic sequencing or particle collision events, most of these actions leave digital traces. As our ability and capacity to measure natural and social phenomena is rapidly increasing at an unprecedented scale, we witness an exponential growth of all these digital traces. This growing digital information is what we call *Big Data*; Data that we generate and acquire far more rapidly than the rate at which we process, analyse and exploit it. Over the last decade, five exabytes (10^{18}) of data were created by humans. Today this amount of information is created in less than two days and is doubling in size every two years [1, 2], generating a new challenge for industries both computationally and storage-wise [3, 4]. While this *Big Data universe* offers unprecedented ways for businesses to gain information to better support their company and serve their customers [5–7], the impact of data abundance extends well beyond business.

In science, the ability to collect and analyse massive amounts of data traces originating from experiments have fuelled numerous advances and unambiguously transformed many research fields from physics and biology [8–10] to computer science and economics [11–13]. But nowhere are these advances more important than in the study of social systems.

As the world is experiencing global crises, demographic explosion and new forms of conflicts, the need to understand social dynamics as well as the structure of social organisations is crucial to resolve the many challenges our society is now facing. Such challenges include global migration, economic instability, epidemic spreading, social divide or organised crime, to cite only a few. The flood of data capturing activities of individuals enables an entirely new scientific approach for social analysis

[14, 15], offering a tremendous opportunity to tackle these challenges. Moreover, the development of computational and storage abilities make it possible for scientists to manage and manipulate such large data to understand, model and predict human behaviour in society.

This thesis aims at illustrating this new approach to analyse social systems in this era of Big Data. While this work provides new understanding arisen from the emergence of large-scale social data , we also demonstrate how this tremendous opportunity can be turned into concrete applications that capture social dynamics in various systems. The contributions in this work evolve around three topics: *human mobility*, *social interactions* and *success*. Even though these topics correspond to different aspects characterising individuals, they are intrinsically related to each other as we will demonstrate in this thesis.

The rest of this introductory chapter is structured as follow: In section 1.1, we will introduce the context in which the different research questions and resulting applications presented in this thesis emerged. In section 1.2 we will give a detailed description of each chapter and finally, in section 1.3, we will provide the list of scientific publications related to the content of this thesis.

1.1 Context of this thesis

This thesis does not address one central question but rather provides instructions on how to turn large-scale data into valuable insights about social dynamics through a collection of concrete projects. These projects investigate three different yet related aspects of human dynamics: *social interactions*, *human mobility* and *individual success*. In this section, we provide details about the context in which this thesis has been done and from which the different questions and projects originated.

Mobile phone data
The first half of the work presented in this thesis relies on mobile phone data obtained through collaborations with telecommunication operators. These data, collected by the network providers for billing purposes, represent a valuable source to capture human activities at a very large scale. As locations of mobile phone towers routing the phone calls are often known, geographical information can be extracted from these data to estimate population activities in space but also to reconstruct movements

of users. These geographical traces have fuelled research on population mapping as well as on human mobility, with profound implications for urban planning [16–18], traffic forecasting [19, 20] or epidemic spreading [21–23]. At the same time, mobile phone data is also found to be a relevant source of social interactions. Indeed, phone calls provide an accurate proxy for social interactions between individuals and have been extensively used by scientists to study communication patterns [24, 25] as well as the configuration of social communities at very large scale [26–28]. Even though these data have been widely used for research on human dynamics, many questions related to human mobility, social interactions and population mapping and in which mobile phone data can provide an answer remain open as we will see.

Spatial mapping of population

Our knowledge of human population numbers and distribution for many areas of the world remains poor [29], despite their importance for policy [30, 31], operational decisions [32] and research [33–35] across many fields. While censuses offer a solution for this data gap, they remain an infrequent and expensive source of detailed population data. Moreover, for many low-income countries the unreliability of estimates, low spatial resolution and complete lack of contemporary data represent further limitations. In collaboration with geographers from the Université Libre de Bruxelles and the University of Southampton, we investigated to what extent could mobile phone data resolve this data gap and provide reliable estimates of population distribution. To address this issue, we designed a cost-efficient method that maps human population at fine spatial and temporal resolutions over large geographical extents, distinguishing ourself from previous research which focussed on urban areas only. Not only can population maps of comparable accuracy to census data and existing downscaling methods in geography be constructed solely from mobile phone data, but these data offer additional benefits in terms of the measurements of population dynamics.

Spatial mapping of communities

By exploiting the communications between mobile phone users, communities of users, i.e. groups of individuals strongly interacting with each other, can also be mapped over large geographical extents. While a large body of research is dedicated to the detection of communities in networks, few studies investigated to impact of regional boundaries defined by governments on the way people interact across space. Yet, understanding and quantifying this impact is fundamental in economic geography and conflict resolution [36–38]. We addressed this issue by

quantifying the correspondence between communities of mobile phone users and regional boundaries in France, finding a remarkable and unexpected relationship between the two. We further assessed the stability of these boundaries by introducing an algorithm that evaluates the spatial sensitivity of users within their corresponding community.

Interplay between social interactions and mobility
While the spatial configuration of communities and their correspondence with administrative regions indicate a strong influence of space on communications, many studies have also pointed out the role space plays on human movements. Even though social interactions and human mobility bears high-level similarities in the role spatial distance plays, they remain as largely separate lines of inquiry, lacking any known connections between the two. Using mobile phone data from three distinct countries, we tested the hypothesis that previously observed spatial dependency captures a convolution of geographical propensity and a popularity based heterogeneity among locations. By separating these factors, we uncovered a scaling relationship between the exponents characterising communication and mobility patterns, allowing us to derive one quantity from the other and hinting for a deeper connection among all systems where space plays a role.

Publication data
The work presented in the second part of this thesis relies on publication data. Over the past few years, we have witnessed an increasing interest for the analysis of publication data, owing partly to the massive growth of global research activity and the recent need to understand the modern scientific enterprise. Publication data not only contain detailed information about research content over time but also allow the study of social interactions throughout co-authorship and citation information, fuelling research on the structure and mechanisms of collaborations [39, 40]. Geographical information can also often be extracted from scientific affiliations present in the data, offering a way to reconstruct career trajectories of scientists and to uncover the role of geography on scientific dynamic [41, 42]. Finally, the citations characterising these data offer clear measures of impact for articles over time and provide a reasonable proxy to uncover the mechanisms underlying scientific success, credit allocation and reputation [43–46]. Despite their recent widespread use in research, the study of such data presents many challenges. Datasets often have a huge size, ranging from thousands to hundreds of millions of publications, and thus require particular storage and computational abilities to access and analyse the data. There are usually no unique identifier

for authors on a publication. A proper disambiguation technique is thus also necessary in order detect the different articles associated to a same individual. Finally, the stored information is usually characterised by many inconsistencies and can cover more than 100 years of data. As a consequence, geographical traces or topics are not always readily available in the data and disambiguation and topic detection techniques are also often required in order to select geographical subsets of authors or articles belonging to a particular subfields.

Quantifying patterns of individual success
We usually tend to think about success as an individual, associating it to novelty or to skills. But success is truly a collective phenomenon influenced by social interactions: for something to be successful, everybody must agree it actually is. While a large body of research in social sciences have been dedicated to the study of human performance in broad domains [47, 48], our quantitative understanding of scientific success is limited. In science, we tend to gauge performance with two distinct measures: productivity, i.e. the number of papers you publish, and impact, i.e. the total number of citations you receive. By taking advantage of publication data where clear measures for productivity and impact of scientist can be extracted, we demonstrated the existence of reproducible productivity patterns leading to the highest impact work of a scientist. We also showed that while highly cited articles appear to be dramatically unpredictable in a scientist career, there are still peculiar statistical features associated to scientist with outstanding publications.

Interplay between success and mobility
If there exists particular patterns of productivity and impact along the career of a scientist, less is known about patterns behind career moves at an institutional level and how these moves affect individual productivity and impact. By taking advantage of the fact that scientists publish somewhat regularly along their career [49, 50], and for each publication, geographical traces can be extracted from their affiliation, individual career trajectories can be reconstructed at a fine scale and in great details. By analysing these trajectories, we found that career movements are not only temporally and spatially localized, but also characterized by a high degree of stratification in institutional ranking. We further showed wether particular movements between specific institutional groups affect or not personal impact.

Collaborations
To conclude, we would like to emphasise that all the work presented in

this thesis results from various international collaborations with several renowned research centers such as the Center for Complex Network Research at Northeastern University, the Orange Labs, the department for applied Mathematics and Computer Science at the Technical University of Denmark, but also the department of Geography and Spatial Ecology at the University of Southampton, the University of Louisville, and the Université Libre de Bruxelles. These various collaborations, bringing actors from Physics, Computer Science, Social Science, Mathematics and Geography, confer a truly interdisciplinary aspect to this thesis and offer valuable insights for individuals involved in a broad range of scientific domains and interested in the study of large-scale social or mobility data.

1.2 Outline of this thesis

Besides this general introduction, this thesis is organised in 7 chapters: one survey chapter and six chapters of scientific contributions. We give hereunder a short description of each of these chapters.

Chapter 2: Analysis of large-scale social data
Recently, we have witnessed a significant increase of interest towards large-scale datasets that capture human activities and its dynamics. As most of our everyday actions are now digitalised and stored, these datasets offer an unprecedented source in both its scope and its scale to study social systems. The nature of these data is various: phone calls, social media posts, emails but also traffic data, online campaigning or publication records, to cite only a few. Geographical information is often present in these data, offering a way for scientists to map human activities in space but also reconstruct career trajectories when time information is also available. These valuable pieces of information enable the development of myriads of applications in broad domains such as epidemic spreading [21–23], urban planning [16–18] or global migration patterns, where space and mobility play a central role. Social interactions can also often be extracted from these data, fuelling applications on social networks [51, 52] and communication patterns [24, 25] but also studies uncovering mechanisms of social influence and success [53–58]. Given the numerous applications and the increasing body of research dedicated to the study of human behaviour, we feel it is important to provide a concise survey of the recent progresses made so far in this area. However, this chapter is not aimed at providing an exhaustive

list of contributions related to the use of large-scale social dynamics data but is rather a selection of major advances relevant to the topics developed in this thesis. As a result this survey will be structure according to the three aspects developed in this work: human mobility, social interactions and the collective phenomenon of success. We will first present major advances on human mobility where geographical traces were exploited to model human movements, distinguishing both micro- and macro-approaches. We will then describe studies that extracted social interactions from these large-scale data and that investigated the role space plays on the way individuals interact and communicate. Along the same line, we will finally survey contributions that looked at social interactions to understand the underlying mechanisms of the collective phenomenon associated to success.

Chapter 3: Dynamic population mapping
This first chapter of scientific contributions is the result of an international collaboration with several geographers from the Université Libre de Bruxelles, the University of Southampton and the University of Louisville and contains the results obtained in a recent journal article [59]. As we show in chapter 2, geographical traces can often be extracted from mobile phone data to map phone call activities in space. In this chapter, we ask to what extent can these phone call activities provide reliable estimates of population densities over large geographical regions. We address this question by designing a method that takes advantage of the correlation between phone calls activities and population numbers to construct population maps. By exploiting mobile phone calls from 2 million users in Portugal, we show that our method produces population estimates of comparable accuracy to census data and state-of-the-art downscaling methods. We then investigate the stability of our approach with respect to the validation procedure and the type of data considered. Finally, we demonstrate that our method can also provide a dynamic population mapping over time to derive mobility information over an entire country. We illustrate this dynamical process with mobile phone data collected from 17 million users from France, showing that our approach offers promising solutions to tackle many challenges in developed but also low-income countries.

Related publication — **Pierre Deville**, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem, "Dynamic population mapping using mobile phone data", in: *Proceedings of the National Academy of Sciences* 111.*45* (2014), pp. 15888–15893

Chapter 4: Spatial distribution of social communities
In this chapter we consider not only geographical traces but also social interactions captured by mobile phone calls between users in France. This information allows us to construct two distinct social networks embedded in space. For the first one, we aggregate the network of users — connected through their phone calls — to a network of towers, where each tower regroups all calls that were passed by users within its vicinity. For the second one, we aggregate the network of users to a network of communes where each commune regroups phone calls passed by users living in that commune. We then describe an algorithm that can efficiently partition these two social networks into almost connected subnetworks called communities. The resulting partitions show that social communities of users in France correspond to geographically cohesive regions strongly influenced by regional borders, suggesting that interpersonal relationships seems to be driven as much by administrative boundaries as geographical proximity. Finally, we investigate the spatial stability of these partitions by introducing a new sensitivity measure that evaluates the attraction between users and their corresponding community. These results are the collection of a journal article [28] and a conference paper [60], both originating from a collaboration with Orange Labs.

Related publication — Vincent Blondel, **Pierre Deville**, Frédéric Morlot, Zbigniew Smoreda, Paul Van Dooren, Cezary Ziemlicki, et al., "Voice on the Border: Do Cellphones Redraw the Maps?", in: *ParisTech Review* (2011)

Chapter 5: Connection between social interactions and mobility
In this chapter, we investigate the role space plays on social interactions and human movements and how these two quantities relate to each other. For this study, we compiled a uniquely rich database consisting of mobile phone data from three distinct countries. As geographical locations of users are known when they perform a phone call, the social fluxes, i.e. number of phone calls, between pairs of locations can be computed. At the same time, movements of users can also be reconstructed from their consecutive calls, allowing us to estimate mobility fluxes as well between pairs of locations. We find that social and mobility fluxes among locations within similar distances follow fat-tailed distributions and that these distributions surprisingly collapse into a single curve once they are rescaled with their average fluxes. This unexpected collapse indicates that the localisation in social communications and human movements

can be decomposed into two independent factors: one that incorporates all the distance dependencies and an independent universal function that captures the inherent popularity-based heterogeneity among different locations. Following this result, we investigate the correlation between social and mobility fluxes and show the existence of a power-law scaling relationship between the two, allowing us to derive one quantity from the other. This relationship together with the observed data collapse allow us to derive a new scaling relationship between critical exponents that characterise spatial dependencies in human mobility and social interactions. Finally, we demonstrate the practical relevance of our results by accurately predicting the population at risk during an epidemic spreading in which mobility fluxes are estimated from social fluxes using our rescaling formula.

Related publication — **Pierre Deville**, Dashun Wang, Chaoming Song, Nathan Eagle, Vincent D Blondel, and Albert-László Barabási, "Scaling Identity in Spatial Networks: Connections between Mobility and Social Interactions", in: (Under review in PNAS)

Chapter 6: Information retrieval in large-scale publication data
Besides phone call data, publication data represent another valuable source of information to study social systems. However, the extraction of meaningful data from publication records is far from trivial as many ambiguities can be present. In this chapter, we present three distinct techniques that aim at resolving these issues. In the first part of this chapter, we address the problem of author name disambiguation. We present an agglomerative method that can automatically merge articles associated to a same author by taking advantage of coauthorship, citation and affiliation information often present in these datasets. We demonstrate the practical accuracy of our method by disambiguating about 1,250,000 scientific articles published in 11 physics journals. Following this method, we introduce another technique to disambiguate geographical traces present in publication data. Using 300,000 affiliations extracted from publication records, we show how this technique can cluster affiliations associated to a same institution with a high precision and recall. We close this chapter by presenting a novel approach to identify publications related to a same particular topic or field. Starting from an initial subset of articles associated to a same field, we show how the citation flow from or to these articles can be used to iteratively detect additional articles belonging to that same field. We illustrate the relevance and accuracy of this approach by identifying all physics articles present in a datasets containing about 40 million publications.

Related publication — Roberta Sinatra, **Pierre Deville**, Dashun Wang, Michael Szell, and Albert-László Barabási, "A Century of Physics", in: (Nature Physics, October 2015)

Chapter 7: Quantifying patterns of scientific success
In science, two measures are often used to evaluate the performance of an individual: productivity, i.e. the number of papers he/she publishes over time and impact. i.e. the number of citations associated to his/her papers . In this study, we investigate the patterns associated to these two individual aspects by exploring thousands of scientist's careers through publication data. We show that productivity gradually improves during a scientist's career and remains steady after the publication of his/her highest impact work. Impact, however, is not characterised by a gradual increase before the publication of highest impact work, nor is more important after. Rather, we show that the probability for a scientist to publish his/her highest impact work is uniform over his entire career, demonstrating the unpredictability associated to highly successful works. We also show that high impact scientists, i.e. individuals who published a very successful work, maintain a career of elevated performance both in terms of productivity and impact. This last observation raises several important questions which we discuss to close this chapter.

Related publication— Roberta Sinatra, Dashun Wang, **Pierre Deville**, Chaoming Song, and Albert-Laszlo Barabasi, "Scientific impact: the story of your big hit", in: (Under review in Science)

Chapter 8: Impact of mobility on scientific success
Affiliations present in publication data constitutes an interesting and valuable information for research. In this chapter, we track affiliation information from publications to reconstruct career trajectories of individual scientists over decades. We show that career movements are common yet infrequent. Most people move only once or twice, and usually in the early stage of their career. We also show that career movements are affected by geography. The distance covered by the move can be approximated with a power law distribution, indicating that most movements are local and moving to faraway locations is less probable. We also observe a high degree of stratification in career movements. People from elite institutions are more likely to move to other elite institutions, whereas people from lower rank institutions are more likely to move to places with similar ranks. We further confirm that the observed stratification is robust against the change in individual performance before

and after the move. When cross-group movement occurs, we show that while going from elite to lower-rank institutions on average results in a modest decrease in scientific impact, transitioning into elite institutions, does not result in gain in impact.

Related publication — **Pierre Deville**, Dashun Wang, Roberta Sinatra, Chaoming Song, Vincent D Blondel, and Albert-László Barabási, "Career on the move: geography, stratification, and scientific impact", in: *Scientific reports* 4 (2014)

1.3 Non-related publications and conference proceedings

We list hereunder the different publications and conference proceedings that are not directly associated to a particular chapter in this thesis. However, these publications are well in line with the subjects developed in this work.

- Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clerot, **Pierre Deville**, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki, "Data for Development: the D4D Challenge on Mobile Phone Data", in: *arXiv preprint arXiv:1210.0137* (2012)
- Vedran Sekara, Roberta Sinatra, **Pierre Deville**, Sebastian Ahnert, Albert-László Barabási, and Sune Lehmann, "The chaperone phenomenon in science", in: (in preparation)

Chapter 2

Analysis of large-scale social data

In this chapter, we introduce some of the most important advances made recently in the study of large-scale datasets that capture human activities. We first review major studies that exploited geographical traces present in these data and which provided the most important advances in the understanding, modelling and prediction of human mobility patterns. We then survey contributions that uncovered the role of space on social communications and that investigated the basic mechanisms of success, both by exploiting social interactions existing in these datasets.

2.1 Introduction

Over the last few years, we have witnessed a remarkable increase in both the scale and scope of social and behavioral data available to researchers. This increasing availability of data capturing activities of individuals enabled an entirely new scientific approach for social analysis [14, 15]. Indeed, with the development of computational and storage abilities, it is now possible for scientists to manage and manipulate such large data to understand, model and predict human behaviour in society in an unprecedented way.

This new *data-driven* approach to study social phenomena through advanced computations is truly interdisciplinary. It involves not only social and behavioural scientists but also computer scientists, mathematicians and physicists. This interdisciplinarity is a key element in its development as traditional tools of social sciences are not sufficient to develop innovative models of the target phenomena. Indeed, unlike biologists or physicists who can measure precisely movements of objects

or behaviours of cells, sociologists often relied on case studies or surveys characterised by incomplete datasets drawn from small samples of the population. These studies, based on stereotypes and averages, were subject to several limitations and potential pitfalls such as their lack of representativeness and independence as well as selection biases [67–70]. The advent of the *digital universe* can remedy theses lack of empirical rigour and shortcomings as data that flows through social media, mobile phone, credit cards, search engines, digital libraries and gadgets offer precise and reliable behavioural information of millions, if not billions, of individuals.

The source and nature of these large-scale behavioural datasets are manifold. Phone logs, for example, remain one of the main used sources of behavioural data in science [71]. Not only it allows us to extract social connections between individuals, but it also contains valuable information about location and time of calls, fuelling research on human mobility [72], urban planning [16–18], spatial communities [26–28], communication patterns [24, 25] or epidemic spreading [21–23]. The emergence of smartphones, which embed a growing set of sensors such as GPS, accelerometers, camera or gyroscope, provides another layer of individual information. Collectively, these sensors are enabling new applications across a wide range of domains such as transportation [19, 20], environmental monitoring [73, 74], social networks [51, 52] or healthcare [75–77]. An other major source of individual and collective information arise from the increasing use of social media. The extraction and analysis of social connections as well as real-time messages and emails from these platforms provide not only meaningful understanding about the emergence of social communities [78, 79] and social contagion [80, 81], but also valuable insights to understand conflicts [82–84], collective response in case of emergencies [85–87] or success dynamic [53]. Over the last decade, the rise of digital libraries has offered another valuable type of data: publication records. Unlike any other sources cited above, data about scientific publications provide reliable information about individual careers but also clear measures of individual performance. This new type of data has fuelled quantitative studies of professional careers uncovering the topology and mechanisms of scientific collaborations [39, 40, 88], performance [54–57] or scientific mobility [42, 89].

Given the amount of research related to the emergence of large-scale social and behavioural data, we feel it is important to survey in this chapter the main scientific contributions related to the topics developed in this thesis. First, in section 2.2 we will present major studies that

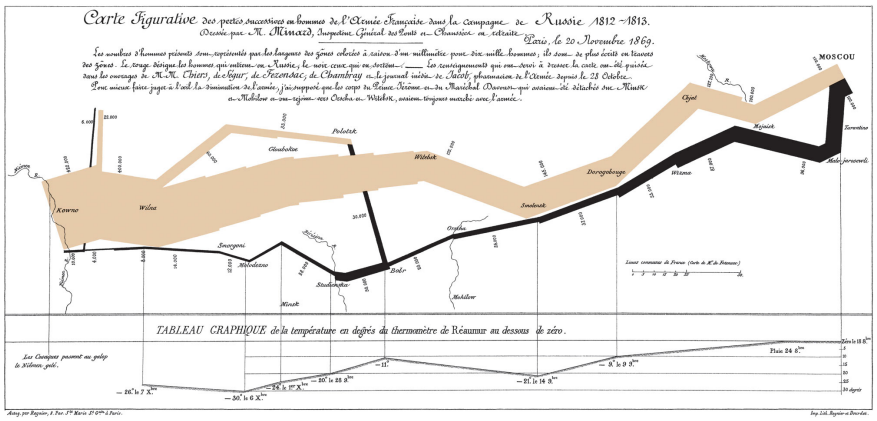


Figure 2.1: **Charles Minard's map of Napoleon's disastrous Russian campaign of 1812.** The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates

exploited geographical and temporal traces to derive mobility patterns of individuals. In section 2.3, we will survey contributions that exploited not only spatial and time information but also social interactions present in large-scale data. In particular, we will first examine research that investigated the role space play on social interactions and then review studies that aimed at uncovering social mechanisms behind success.

2.2 Human mobility

Capturing mobility patterns of individuals has alway fascinated scientists. One of the earliest example is certainly the graphic produced by the engineer and statistician Charles Joseph Minard in the beginning of the 19^{th} century and which depicted the flow of french soldiers in space during the disastrous Russian campaign of 1812 (Fig. 2.1). This first quantitative analysis of population flow remains one of the most famous study involving numerical mobility data. By meticulously collecting information about the positions of soldiers over time, Minard produced a compelling map that convey more information than any written work or paintings at that time and which perfectly illustrate the futility of Napoleon's attempt to invade Russia.

Since then, the understanding of the basic laws that govern population flow and individual mobility remains a central issue for scientists. Indeed, mobility has been proven to be a key driving force in various domains such as urban planning [16, 90], traffic forecasting [91] and human migration [92] but also in crucial spatiotemporal phenomena such as the spread of biological [93, 94] or mobile viruses [22, 95]. In this section, we will survey some of the main contributions and discoveries related to human mobility and which took advantage of large-scale spatiotemporal data. These contributions will be depicted under three different subsections: (i) micro-approaches, which focus on the trajectory of an individual, (ii) macro-approaches, which are concerned with explaining aggregate migration patterns and finally (iii) applications highlighting the importance associated to the understanding of human mobility.

Individual mobility patterns

As humans travel on many spatial scales over short periods of time, direct quantitative assessments of individual human trajectories has always been difficult in the past. The increasing availability of large-scale spatiotemporal information, however, offers a way to address this issue. A decade ago, Brockmann *et al* were among the first to exploit geographical traces at a very large-scale. By collecting data at online bill-tracking systems, the authors were able to analyse individual trajectories of about half a million bank notes to infer the statistical properties of human dispersal [96]. As bank notes are carried by individuals, their dispersal offers a good proxy for human movements. The authors suggested that these movements were characterised by Lévy statistics and were best modelled by a continuous-time random walk (CTRW) with fat-tailed displacements and waiting-time distributions, corroborating the observation that people tend to travel mostly over short distances and occasionally over longer ones.

However, as pointed out by Gonzáles *et al* soon after, bank notes movements reflect composite motions of more than one individual. As bills are passed from one person to another, they do not quite capture individual trajectories. Using mobile phone data, Gonzáles *et al* were able to correct this issue by reconstructing time-resolved trajectories of about 100,000 mobile phone users. While they observed a fat-tailed distribution of human displacements with power-law decay, corroborating the observations of Brockmann *et al*, they showed that individual mobility exhibits a high degree of temporal and spatial regularities as users tend to return to a few highly frequented locations [97], contrasting with the Lévy/CTRW framework previously proposed. More particu-

larly, they showed that human motions, when rescaled and aligned along their principal axes, are very similar among individuals and, as a result, that human mobility can be captured by a single function. Song *et al* supported these findings by showing that human traces are barely random and that the results of the Lévy/CTRW models are in systematic contradiction with empirical results [98]. As an alternative, they introduced a new model that incorporates exploration and preferential return mechanisms, both unique to human mobility but missing in the random-walk model. In their model, each time a user decides to move, he/she is facing two different choices: (i) explore an additional location not yet visited with a probability decreasing in time or (ii) return to a location he/she already visited with a probability proportional to the number of visits the user previously had to that location. This simple yet self-consistent conceptual framework offers an improvement over previous modelling efforts. Not only it accounts for the uneven probability of a user to visit a given location over time but it also captures the decreasing tendency to visit previously unvisited locations and explains the ultra-slow diffusion associated to the mobility process.

Even though human activity patterns might appeared to be random, the recent modelling frameworks suggest that our movements are characterised by a high degree of temporal and spatial regularities. This raises an important question: To what extent are our individual movements predictable ? To address this question, Song *et al* looked at 50,000 individual trajectories extracted from mobile phone data and quantified the interplay between the observed regularities and the randomness of movements [99]. To do so, they measured the entropy associated to each individual's mobility pattern, which captures the degree of predictability of a user's trajectory. They found that the uncertainty in a typical user's whereabout is drastically reduced by the spatio-temporal correlation present in the user's mobility pattern and that users trajectories hide an unexpectedly high degree of potential predictability. As the distribution of distances over which users travel on a regular basis are fat-tailed and thus highly heterogeneous, one would expect to observe that same characteristic for the predictability. Yet, the authors observed that the degree of predictability remains constant across the population, no matter the typical distances covered by users. These results were later supported by Lu *et al* who also observed a high predictability of human movements even under much more extreme conditions and in low-income settings [100].

Aggregate mobility patterns

While much research focused on the understanding, the modelling and the prediction of individual human trajectories, numerous advances related to aggregate migration patterns of the population have also been made. Unlike studies on individual trajectories, research on aggregate mobility data focuses on the flow of individuals from one point to another in a city or a country.

Intuitively, it would seem reasonable to think that the number of movements between two locations depends on the number of possible contacts between the two populations characterising both places and on the distance between them, or more generally on the cost of travelling from one to the other. One of the first framework that incorporates these two aspects are *Gravity* models. Introduced over 60 years ago in its contemporary form by Zipf [101], this type of model considers the number of people T_{ij} that moves in between place i and j to be proportional to some power of the population of the origin (p_i) and the destination (p_j) and to decay with the distance between them ($r_{i,j}$) as

$$T_{i,j} \sim \frac{p_i^{\alpha} p_j^{\beta}}{f(r_{ij})} \tag{2.1}$$

where α and β are adjustable exponents and the deterrence function $f(r_{ij})$ describes the effect of space and depends on the system studied. This framework has been shown to predict fluxes in numerous situations. Balcan *et al*, for example, applied this framework to model the flow of commuters at a global scale [102]. By analysing short-scale commuting flows and long-range airline traffic data, they were able to provide a global description of commuting patterns which was best-fitted by Eq. 2.1 with sets of exponents varying according to the distance considered. Other studies include the one from Jung *et al* who also used one of the form of the gravity model to describe traffic flows on Korean highways [103] or, similarly, the study of De Montis *et al* who looked at inter-urban traffic flows in the region of Sardinia [104]. Other analyses have also shown that these types of models can not only predict human mobility fluxes but can also capture other kinds of flows such as bank notes movements [105], cargo ship movements [106] or even inter-urban communication flows [24, 107].

Despite the predictive success of gravity models in various studies, this framework met with considerable objection over the years. One of the main point of discussion is the use of the euclidean distance in the deterrence function of Eq. 2.1, which suggests that human movements follow some universal law constrained by the euclidean distance. Indeed,

some scientists suggested that there is no direct relationship between mobility and distance, and that distance is a surrogate for the effect of *intervening opportunities*. First introduced by Stouffer [108], this competing theory suggests that the number of movements from an origin to a destination do not depend on the distance itself but rather depends on the number of opportunities closer than the destination. Displacements are thus driven by the accessibility of resources satisfying the objective of the trip rather than by the distance as in gravity models. This competing framework has successfully been used in numerous studies to explain movements of various nature such as migrations [109], work-trip [110], but also telecommunication flow [111].

More recently, Noulas *et al* identified a new universal law for urban mobility derived from the concept of intervening opportunities [112]. By extracting about 35 million human movements within several cities from a location-based social network, they highlighted the decisive role of spatial density of places, i.e. opportunities, on urban movements. By deriving a rank-distance measure that incorporates the spatial distribution of opportunities, they were able to accurately capture human displacement patterns in different urban environments, correcting the observed predictive discrepancies of the gravity models across different cities [113].

The same year, Simini *et al* proposed a new model to predict migration patterns, supporting the results from Noulas *et al*. Based on radiation and absorption processes, their model not only incorporates the spatial distribution of population to correct for the predictive discrepancies of gravity models but is also parameter-free [114]. This lack of parameters corrects numerous other limitations associated to Eq. 2.1 such as its lack of theoretical guidance for the functional form of $f(r)$ which can vary according to the system studied or the need for training data to fit the different parameters $[\alpha, \beta, ...]$. Moreover, the model accurately captures mobility and transportation flows from distinct socio-economic phenomena (hourly travel patterns, migrations, communication patterns and commodity flows), demonstrating its generality of use on a wide range of timescales and places.

Applications

As described in the previous paragraphs, numerous researches provided valuable insights on the way an individual or a group of individuals moves across space. Beyond the understanding, several studies aimed at modelling human mobility, documenting the mechanisms and pre-

dictability of individual movements as well as offering predictive tools to accurately derive population flow between locations. While these major advances might answer several important sociological questions, one can wonder what applications these spatiotemporal data and modelling efforts are good for.

The importance of space and mobility patterns appears clearly in the study of epidemic spreading. In pre-industrial times, human movements were limited to relatively short distances over time as few travelling means were available. As a consequence, the spread of disease was thus mainly a spatial diffusion phenomenon, as confirmed by historical studies [115], and their propagation could easily be described with simple scheme [116]. However, our modern societies are sharply contrasting with this simple picture. As humans now travel on many spatial scales over short periods of time, epidemiology is growing complex and cannot be described by pure spatial diffusion. As demonstrated for the SARS epidemic in 2003 [117] or for the recent Ebola outbreak in 2014 [118], modern epidemics can now quickly propagate to far-reached regions due to fast evolving transportation systems and not only diffuse out near their origin. Hence, estimates on population mapping, transportation fluxes, population movements on a local and global scale are key ingredients in current epidemic modelling efforts [102, 116, 119–125] and human mobility research remains crucial for our understanding of past and future epidemic behaviour and our capabilities to predict their evolution.

Studies on human mobility also provide valuable insights for research on urban planning and traffic forecasting. By using electronic ticketing system data from London as a proxy for human movements, Roth *et al* [126] investigated the structure and organization of the city at an unprecedented scale. They not only identified and quantified polycentricity characteristics of London but also unpacked the complex patterns of flow in the different subcenters, shedding new light on the dense structure of its centres and providing an initial approach to modelling flows in urban systems. In an other study, Calabrese *et al* [127] derived mobility movements from mobile phone data to understand the correlation between social events people go to in Boston and their home location. They found that the type of an event is strongly correlated with and can be predicted from the distributions of origins of people attending that event, providing crucial insights for city management and traffic congestion.

As mobility traces can be readily monitored to offer temporal population estimates in space [128], various applications relying on population mapping also benefited from advances on human mobility. The analysis of tourist activity for example can provide valuable information to assess local economy of places in time [129–131]. Other applications includes the estimation of distribution of visitors from mobile traces to assess the evolution of the attractiveness of points of interests in an area [132] or the collection of spatiotemporal and georeferenced social media data, such as pictures, to model and to improve knowledge about geographical areas and to detect temporal trends [133].

2.3 When interactions come into play

As we have seen in the previous section, spatiotemporal data provides valuable information to derive human mobility laws, fuelling myriads of applications on social systems. Yet, another type of information remains important (if not the most important) to understand the complex nature of social systems: *interactions*. From professional, friendship or family ties, to communications, collaborations or citations, interactions are the fabric of our society. Many studies have highlighted the fundamental role they play in the understanding, modelling and prediction of social dynamical systems, leveraged by the emergence of network science tools [134]. As research on social interactions is extremely vast, we will focus here on two important aspects relevant to this thesis: (i) the relationship between space and social interactions and (ii) the social mechanisms behind success.

Interplay between space and social interactions
We have shown in the previous section that space plays an important role on the way individuals move. As in general human tends to minimise their effort [135], it is also reasonable to think that the influence of space would be reflected on the way we create and maintain social ties through our social interactions.

As a proxy for social interactions, Lambiotte *et al* extracted 810 million communications (phone and text messages) between 2.5 millions users in a mobile phone dataset. As geographical home locations of users were known, the authors were able to derive the distance associated to each communication. They observed that the probability $P(d)$ to have two interacting individuals separated by an euclidean distance d decays

as $P \sim 1/d^2$, in line with the idea of the gravity law developed in Eq. 2.1. This result was later supported by Krings *et al* who measured the intensity of mobile phone communications between cities and obtained results consistent with this previous observation.

This decay with distance is not only present in mobile phone communications but in many others social systems. A decade ago, Liben-Nowell *et al* analysed the social network of bloggers in the US and observed a spatial scaling of $P \sim 1/d$ for the probability to have friends at euclidean distance d [136]. This spatial scaling was later confirmed in numerous studies such as for the density of social network contacts on Facebook or for volumes of email traffic between cities [137, 138].

Even though the observed spatial scalings may differ depending on the social system investigated, the influence of distance on the way people interact is rather clear as most studies suggest that our friends tend to be located nearby. Yet, this interplay with space goes beyond friendship. Recently, Pan *et al* showed that geography was also influencing the dynamic of science [41]. By assigning scientific articles to geographical locations, the authors were able to study the effect of distance on citations but also scientific collaborations. They observed a similar trend than for friendship as the strength of flow of citations and collaborations between cities decreases with distance and follows a gravity law.

This strong relationship between social interactions and geography led scientists to wonder if one could predict one from the other. By integrating this interplay in a maximum-likelyhood approach, Backstrom *et al* proposed an algorithm that predicts the physical location of social-media users based on the location of his/her friends, obtaining a better accuracy that standard IP-based methods [138]. Conversely, Crandall *et al* as well as Wang *et al* investigated the ability of geography and mobility information to infer social ties between people. These studies showed not only that co-occurrences in space of individuals and their social ties are strongly correlated but also that mobility measure alone yield surprising predictive power [139, 140].

Success as a collective phenomena

We usually tend to think about success as an individual, associating it to novelty or to skills. But success is truly a collective phenomenon influenced by social interactions. From the size of your audience as an artist, the number of votes as a politician, the number of citations as a scientist or total number of views as a youtube video, success truly

depends on the collective perception people have of you or your work: for something to be successful, everybody must agree it actually is. If we accept the collective nature of success, its signature and mechanism can then be uncovered by studying social systems through the large amount of data now available.

One of the main features of successes in our current society such as hit-songs, best-selling books or nobel papers is the order of magnitude there is between their success and the average [141–143]. Social scientists often attribute this inequality in cultural markets to social influence. Indeed, as individuals are now facing an overwhelming number of choices in these markets, with almost no information about them, their choices are likely to settle on already successful items [144], mimicking but also trusting choices made by other individuals [145]. This effect can also often be reinforced by structural characteristics in these systems; populars book are likely to be more visible in a library, famous politicians will appear more often on the news or highly cited papers will appear on top of google scholar search results. Both these individual and structural factors tend to make successful objects or people even more successful, leading to a cumulative advantage and increasing the disparities between them.

A large body of research has been dedicated to the understanding of this cumulative advantage present in many systems. This characteristic has been studied in a variety of fields and under a variety of names: "increasing return" in economics [146], "Matthew effect" and "success-breeds-success" in sociology [147], or "preferential attachment" in complex networks [148]. One of the major large-scale studies that suggested the existence of this effect in many complex systems was produced by Barabási *et al.* By analysing large databases describing the topology of networks in a wide range of fields such as the World Wide Web or citation patterns, the authors observed that all of them where characterised by a scale-free state, i.e. the probability for a element in these systems to be connected with k other elements decays as a power-law [148]. To explain this scaling characteristic, the authors successfully introduced a model that incorporates the notion of cumulative advantage: when a new element enters the system, it is more likely to connect to popular elements, i.e. elements that already have many connections. This observation has since been shown to characterise a wide range of areas, offering tangible proofs for the origin of the social and economical disparities present in many competitive systems [149–151].

This notion of cumulative advantage was later investigated by Petersen *et al* who introduced a model that quantitatively incorporates cumulative advantage mechanisms to describe career development. They validated their model on the careers of 400,000 scientists extracted from six different journals and 20,000 athletes in four sport leagues, demonstrating that longevity and past success of an individual lead to a cumulative advantage in further developing his or her career. Van de Rijt *et al* also supported this cumulative advantage effect by quantifying the "success-breeds-success" dynamic in different reward systems. To do so, they constructed an experiment on four naturally occurring systems representing distinct form of personal success: endorsement, financial gain, social status, and social support. In their experiment, they randomly selected individuals and gave them an early artificial contribution of success. In the four distinct systems, they observed a significant increase in rewards for individuals who received an early contribution. But at the same time, they also observed a decreasing marginal return of success, suggesting that this noticeable social feedback might be bounded.

While these studies have been successful in understanding the underlying mechanisms of success and popularity dynamics, most of them lack any predictive power for the success dynamic of individual items. As suggested by Watts *et al*, this inability to predict individual outcomes does not necessarily come from a lack of competence, but could simply originate from the inherent impossibility of the task [58]. To support this theory, the authors constructed a website where people could choose, listen and rate musics and where the social influence could be controlled. By analysing the behaviour of nearly 30,000 users they found support for their idea and showed that social influence at a micro level led to high unpredictability at macro level, which could explain the difficulty to capture future success or popularity of individual items in many systems.

Despite these limitations, we have witnessed over the last few years a rapid escalation of interest towards individual success, fuelled by the increasing availability of finer-grained data. While scientists usually investigated the phenomenon with respect to their own field, it now engages social scientists, computer scientists, economists, physicists, and mathematicians alike in a large variety of area.

This revival of interest is particularly pronounced for research that focuses on the understanding of the scientific enterprise. As the success of an article (or a scientist) is often measure through its impact, i.e.

the number of citations it acquires over time, the emergence of accurate individual data from digital libraries as well as the increasing computational capabilities have recently offered an unprecedented opportunity for scientists to understand the mechanisms of success for individuals or individual papers.

By extracting accurate data about 400,000 scientific articles from 1883 to 2012 as well as their citations over time, Wang *et al* recently developed a mechanistic model that for the first time accurately captures the citation dynamics of papers. Not only their model uncovers the basic mechanisms that govern scientific impact, but it also predicts accurately the temporal success of individual papers.

A few months later, Servia-Rodriguez *et al* investigated scientific success evolution with respect to co-authorship. Interestingly, the observed a strong interplay between the two: successful scientists tend to play a central role in the co-authorship network, supporting the idea that collaborations play a fundamental role on individual success. At the same time, Sarigol *et al* also investigated the interplay between success of individual papers and their social dimension captured by collaborations. Not only they also observed a strong dependence between the two aspects, but they also highlighted the power of collaborations to predict future individual success.

Part I

Social interactions and human mobility

Chapter 3

Dynamic population mapping

In this chapter, we address the issue of dynamic population mapping by investigating the extent to which phone call activities provide reliable estimates of population densities. First, we introduce the mobile phone datasets that are used to derive the spatial activities of individuals. Second, we present a method that can cost-effectively provide accurate and detailed maps of population distributions over large geographical regions based on the collected mobile phone data. Third, we evaluate the stability of the method and quantify its robustness as well as its extrapolation capacity to map population in data-scarce countries. Finally, we show the potential of our approach to derive population movements based on temporal variations of population densities and discuss its applications in low-income countries. This chapter is largely based on the research of *P. Deville et al* [59].

3.1 Introduction

Our knowledge of human population numbers and distribution for many areas of the world remains poor [29], despite their importance for policy [30, 31], operational decisions [32] and research [33–35] across many fields. In the 1990s, a growing interest in the global mapping of human populations emerged [152, 153], leading to the advanced development of methodologies that undertake the spatial downscaling of human population count data from censuses summarized over large and irregular administrative units to grid squares of 100m-5km resolution [154–160]. Initial efforts to downscale these data used simple areal weighting methods [154, 161] or dasymetric modeling approaches [157–159], which use ancil-

lary layers to redistribute population counts within administrative units [162]. Modeling techniques that spatially downscale population numbers into gridded datasets continue to be refined, with basic dasymetric models increasing in sophistication, incorporating multi-scale remotely-sensed and geospatial data and making improvements in the type of statistical algorithms used in the modeling process [163–165]. These detailed population databases have been proven to be crucial for studies reliant on information about human population distributions, typically for calculating populations at risk of human or natural disasters [166–168], to assess vulnerabilities [35, 169] or to derive health and development indicators [31, 33, 170, 171]. However, despite improvements these data still have many limitations.

Regardless of how sophisticated these methods are, they remain largely constrained by population count data from censuses that form the basis for the estimation of population distributions across large areas [154–161]. While increasing usage of global positioning and geographical information system technologies has supported the improved collection of census data and their processing, censuses remain an infrequent and expensive source of detailed population data. Moreover, for many low-income countries, the unreliability of estimates, low spatial resolution and complete lack of contemporary data represent further limitations. These restrictions mean that the latest health indicators or estimates of populations at risk can often be based on outdated and coarse input population data [170, 172, 173], a particularly restrictive feature when accurate contemporary numbers may be required for disaster impact assessments, epidemic modeling or conflict relief planning. Human populations are dynamic, moving daily, seasonally and annually, resulting in rapidly changing densities. Attempts have been made to model and map these dynamics for high-income countries [164, 174], but the data streams upon which such models are based are currently unavailable for most of the world, particularly resource poor regions.

In this chapter, we show how the proliferation of mobile phones (MPs) offers an unprecedented solution to this data gap. The global MP penetration rate (i.e. the percentage of active MP subscriptions within the population) reached 96% in 2014 [175]. In developed countries, the number of MP subscribers has surpassed the total population, with a penetration rate now reaching 121%, while in developing countries it is as high as 90%, and continuing to rise [175]. These data provided by communication tools are opening up new opportunities for studying socio-spatial behaviors [14, 65, 176, 177]. MP call detail records have

been used in the past for studying human mobility patterns at the individual level [97, 99, 178] or for mapping human movements and activities using aggregated data [16–18, 179, 180]. Most of these studies have focused on specific cities or city neighborhoods or groups, and were aimed at understanding traffic flows [179], mapping the intensity of human activities at different times [17, 18, 180] or exploring seasonality in foreign tourist numbers and destinations [129, 130]. Population movement analyses based on MP data are particularly promising for improving responses to disasters [100, 181] and for planning malaria elimination strategies [125, 182, 183]. However, to date, these data have not been assessed in their capacity to map human population at fine spatial and temporal resolutions over large geographical extents.

Using Portugal and France as case studies, we examine how aggregated MP data could be used efficiently to map population distributions at the country scale. We then assess how such predictions compare to existing state-of-the-art downscaling methods. Next, we quantify the robustness of our methodology over social groups and regions and its extrapolation capacity in order to facilitate widespread use. We then show that this approach can reveal unmeasurable patterns in space and time. Finally, we end this chapter with a discussion on how our methodology relates to privacy and data access concerns.

3.2 Data description

This section describes the two main types of data sources exploited in this chapter. First, we present the mobile phone datasets used to derive population density estimates in two distinct countries. Second, we describe census data which provide a baseline to calibrate the method developed in the chapter as well as to assess its accuracy.

Mobile phone data MP networks are composed of cells, i.e. geographic zones around a mobile phone tower (Fig 3.1). Each MP communication can be located by identifying the geographic coordinates of its transmitting tower and the associated cell. This network-based positioning method is simple to implement and its accuracy directly depends upon the network structure; the higher the density of towers, the higher the precision of the MP communication geo-localization [184]. Records detailing the time and associated cell of calls from anonymous

users therefore provide a valuable indicator of human presence. Two large datasets of MP calls obtained from major operators in Portugal and France were used as proxies for population activity in the countries. The two datasets cover the following periods: July 2006 to August 2006 and November 2006 to June 2007 (10 months) for Portugal and May 2005 to October 2007 (5 months) for France. Both datasets contain more than a billion calls from 2 million users in Portugal (20% of the total population) and 17 million users in France (30% of the total population). According to the operators, their penetration rates were uniform over the country at that time. Only calls were considered; text messages were excluded. Phone calls corresponding to MP contracts from companies were excluded as well from both datasets in order to include only MP contracts of individuals. For each call, the originating and receiving towers and the day when the call was made were obtained. The time when the call was made and a user identifier were available for Portugal only. Phone call data from Portugal are used to illustrate the accuracy, overall stability and the dynamic aspect of the presented method while data from France are only used to assess its extrapolation capacity and its dynamic aspect as well.

Census data Census population data were obtained from the National Institute of Statistics of Portugal for the year 2011 [185] and from the National Institute of Statistics and Economic Studies of France for the year 2007 [186]. Census population data were matched to administrative units using identifier codes. For both countries, the finest available administrative unit level was used (ADM-5), which corresponds to *Freguesias* in Portugal (n = 2,882) and *Communes* in France (n = 36,610). The spatial resolution of administrative units is similar in France and Portugal, with average spatial resolutions (i.e. square root of the land area divided by the number of administrative units) of 3.9 km and 5.6 km, respectively.

3.3 Population mapping methods

In this section, we introduce a mapping method based on mobile phone data (MP method). To evaluate its accuracy, we present a state-of-the-art mapping method which is based on a dasymetric modeling approach and incorporates a wide range of remotely-sensed and geospatial data (RS method). We then discuss and compare the accuracy of both meth-

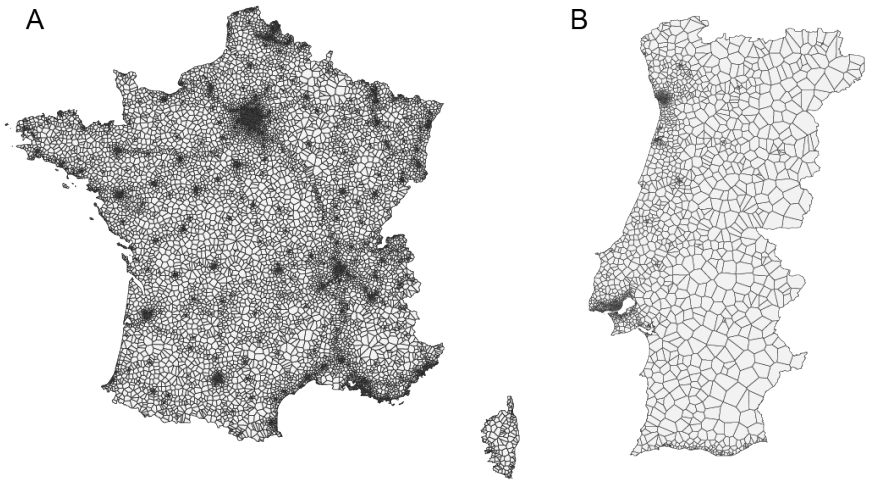


Figure 3.1: **Mobile phone tower maps.** Map of Voronoi cells corresponding to mobile phone towers in **(A)** France and **(B)** Portugal

ods as well as introduce a third method that combines the two different approaches.

Population mapping using mobile phone data
As described previously, MP networks are composed of cells, i.e. geographic zones around a mobile phone tower. For each tower j, the total number of different users T_j that made or received phone calls from/to that tower is known. When one makes a phone call, the network usually identifies nearby towers and connects with the closest one [187]. The coverage area of a tower j can thus be approximated using a Voronoi-like tessellation [188] (Fig 3.1). The Voronoi polygon associated with tower j is denoted v_j. The MP user density of the polygon v_j, denoted as σ_{v_j}, is then equal to T_j/A_{v_j} where A_{v_j} is the area of the Voronoi polygon corresponding to tower j.

The estimation of the population density for an administrative unit c_i based on the MP user density σ_{v_j} is a two-step method. First, the MP user density σ_{c_i} for c_i is computed with the following equation:

$$\sigma_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} \sigma_{v_j} A_{c_i \cap v_j} \tag{3.1}$$

where A_{c_i} is the area of administrative unit c_i and $A_{c_i \cap v_j}$ is the intersection area of c_i and the Voronoi polygon v_j. An illustration of how

the MP user density is estimated for an administrative unit is given in Figure 3.2.A. Secondly, MP user density values σ_{c_i} assigned to each administrative unit are compared with baseline census-derived population densities available in a training set, denoted as ρ_{c_i}. The approach is modelled as follows:

$$\rho_c = \alpha\sigma_c^\beta \tag{3.2}$$

where $\rho_c = [\rho_{c_1}, \rho_{c_2}, \ldots, \rho_{c_n}]$ and $\sigma_c = [\sigma_{c_1}, \sigma_{c_2}, \ldots, \sigma_{c_m}]$. The parameter α represents the scale ratio and β the super linear effect of population density ρ_c on the MP user density σ_c. This can be transformed to $log(\rho_c) = log(\alpha) + \beta log(\sigma_c)$, where a standard linear regression model with population-weighted least squares is applied to estimate the two parameters α and β. Population density estimates $\hat{\rho_c}$ of all administrative units can then be derived using Eq. 3.2. The population densities $\hat{\rho_c}$ are then adjusted to make the total estimated population, $\hat{P} = \sum_i \hat{\rho}_{c_i}$, match the census-derived national population P:

$$\hat{\rho_c} = \frac{P}{\hat{P}}\alpha\sigma_c^\beta \tag{3.3}$$

While the density of MP users is chosen here as input to estimate the population density in an area, it is important to note that other quantities such as the MP call density can be used as well. The precision and stability impact of choosing the MP user density over the MP call density over different time windows is discussed in the next section.

Population mapping using landcover and geospatial data
Recently, a new modelling method has been developed by the WorldPop project [189]. The aim of this method is to generate gridded predictions, i.e. pixels, of population density at $\approx 100m$ spatial resolution. To do so, two types of spatial information are used in input: (i) land cover data which categorise the type of landscape, e.g. urban, industrial, herbaceous, water areas or even elevation, and (ii) geospatial data such as distance to roads, villages, railways, schools or hospitals. A machine learning technique, incorporating these spatial variables as well as a training set of census data, is then used to associate an estimated population weight to each pixel. An illustration of such a weight layer is given in Fig. 3.2B. Given this estimated weight layer, the total population in the country is then redistributed into pixels according to their associated weight as follow

$$\hat{\rho}_i^{RS} = \frac{w_i}{\sum_j w_j}\frac{P}{A_i} \tag{3.4}$$

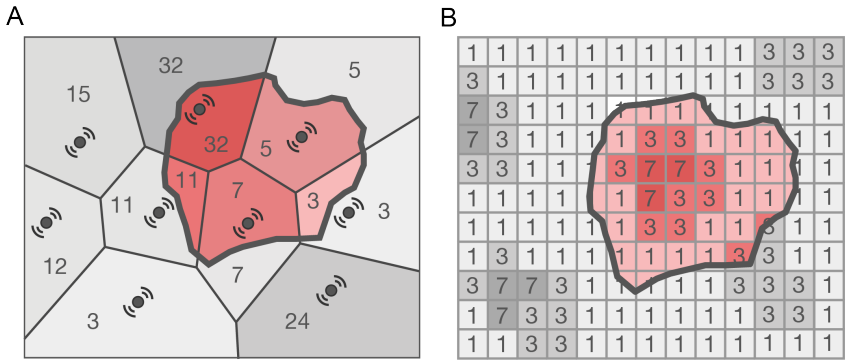


Figure 3.2: **Illustration of MP and RS methods (A)** Illustration of the MP method, where Voronoi polygons are built based on the spatial configuration of MP towers. The MP user density (or call density) of an area (red polygon) is derived from the proportion of Voronoi polygons intersecting that area, as described in Eq. 3.1. **(B)** Illustration of the RS method, where a relative weight is assigned to each pixel according to its environmental and infrastructural characteristics. The estimated population density of a commune (red polygon) is given by the average population density of pixels that fall within the commune.

where $\hat{\rho}_i^{RS}$ is the population density in pixel i, w_i is the weight associated to pixel i, P is the total population in the country and A_i is the area associated to pixel i. As illustrated in Figure 3.2B, the estimated population density of a commune (red polygon) is then simply given by the average population density of pixels $\hat{\rho}^{RS}$ that fall within the commune.

Comparison of mapping accuracy

The ability of the MP data-based approach to accurately downscale census population data is compared to that of the RS method for Portugal. For comparative purposes, the same spatially-stratified training dataset (Norte region) was used to estimate nighttime population densities using both the MP and RS methods. Fig. 3.3 shows the nighttime maps produced for Portugal using the MP (Fig. 3.3B,E) and the RS methods (Fig. 3.3C,F), compared to baseline census-derived population densities (Fig. 3.3A,D). At the national scale, both methods show similar spatial patterns that match baseline data, with major cities being clearly identifiable (Fig. 3.3A-C). However, the close-up on the capital city of Lisbon highlights clear differences in estimated population densities visible at finer spatial scales (Fig. 3.3D-F). The spatial detail of the MP method relies on the density of towers, which is substantially higher in urban ar-

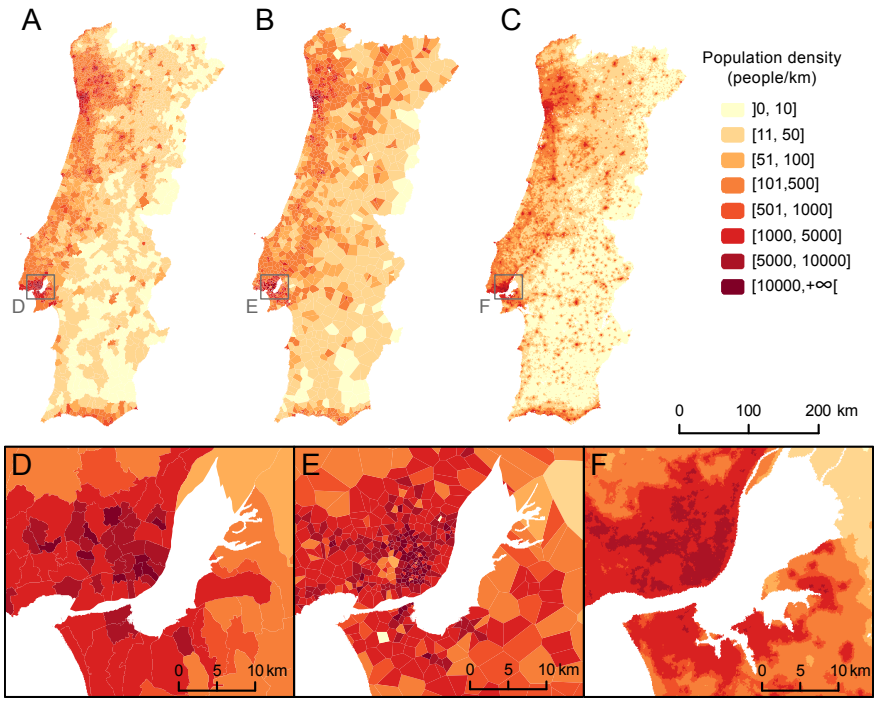


Figure 3.3: **Comparison of predicted population density datasets with baseline data for mainland Portugal. (A)** Population density as calculated from the national census at Administrative unit level 5 (ADM-5; freguesia), **(B)** population density at the level of Voronoi polygons, as estimated by the MP method, and **(C)** population density at the level of 100 x 100 m grid squares, as estimated by the RS method. **(D)-(F)** show close-ups around the capital city Lisbon.

eas, whereas the spatial detail of the RS method depends on the spatial resolution of the geospatial datasets used in the mapping process, which often do not capture intra-urban variations.

Precision and accuracy statistics, including the Pearson product-moment correlation coefficient (r) and root mean squared error (RMSE), are calculated to compare the performance of the MP and RS downscaling methods, using the baseline census-derived population densities as reference (Fig. 3.4). The wider cloud observed for the MP method (Fig. 3.4A) indicates a lower precision, especially in low-density areas, and the RS method tends to overestimate population densities in low-density areas and underestimate in high-density areas (Fig. 3.4B). Globally, the RS method is found to be more precise than the MP method

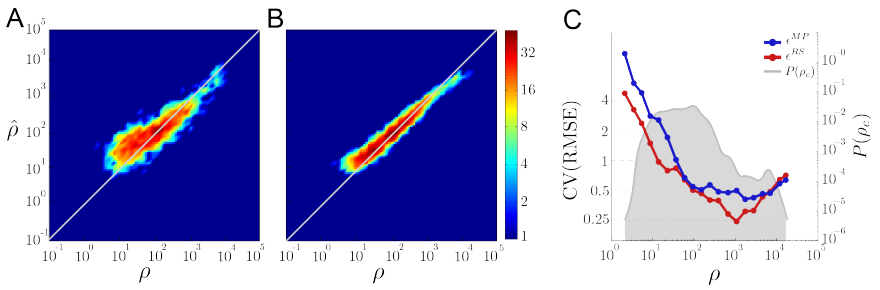


Figure 3.4: **Accuracy of MP and RS methods.** Relation between baseline and estimated population densities using **(A)** the MP method and **(B)** the RS method. **(C)** RMSEs normalized by the average population density of intervals, for the MP (blue) and RS (red) methods, on a logarithmic scale. The shaded area represents the absolute population count per interval. Both methods were calibrated on the Norte region ($n = 1,425$) and their accuracy was assessed on the rest of the country ($n = 1,457$)

($r_{MP} = 0.89$; $r_{RS} = 0.92$). Fig. 3.4C shows how the normalized RMSE of both methods decreases with population density. RMSE values are always higher for the MP than the RS method, except in high-density areas. Overall however, the MP method is found to be slightly more accurate than the RS method ($RMSE_{MP} = 796$; $RMSE_{RS} = 850$), given the importance of densely-populated areas in the RMSE calculation.

Combination of MP and RS methods

In order to optimize both spatial and temporal resolutions, the MP method can be combined with the RS approach described above. In a first step, we estimate the nighttime population of each Voronoi polygon v_j that corresponds to the coverage area of tower j using the MP method. Then, the population of v_j is disaggregated to $\approx 100m$ grid squares using the Random Forest approach developed for the RS method. The combination of both methods (COMB) captures the spatial details resulting from the RS method, especially in more rural areas where the density of MP towers is low. It also captures the spatial details resulting from the MP method, especially in urban areas where the distance between MP towers is often finer than the spatial resolution of the geospatial datasets used in the RS method (Fig. 3.5). Here we used the same training (Norte region) and evaluation datasets as in Figure 3.3 and extracted accuracy statistics. An overall higher accuracy is achieved with the COMB method compared to the MP and RS methods ($RMSE_{MP} = 796$; $RMSE_{RS} = 850$ and $RMSE_{COMB} = 684$), while

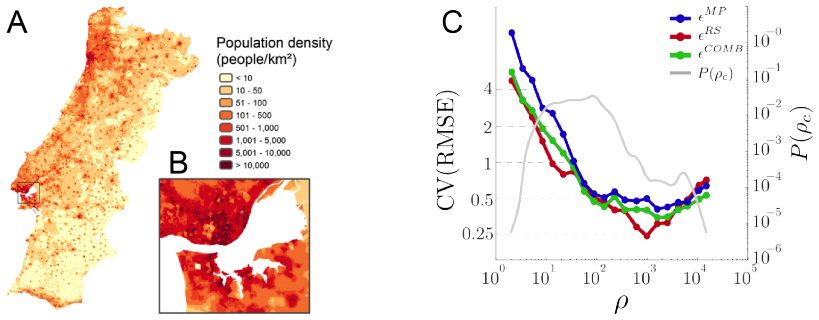


Figure 3.5: **Accuracy of the COMB method.** Population density at 100 x 100 m spatial resolution, as estimated by the combination of the MP and RS methods: **(A)** mainland Portugal with **(B)** close-up around the capital city Lisbon. **(C)** RMSEs normalized by the average population density of intervals, for the MP (blue), RS (red) and COMB methods (green). To aid visualisation, RMSEs are plotted on a logarithmic scale. The grey line represents the absolute population summed by population density intervals.

the overall precision is identical to the MP method but lower than the RS method ($r_{MP} = 0.89$, $r_{RS} = 0.92$ and $r_{COMB} = 0.89$). Even though the RMSE is lower than the RS and MP methods for the COMB method in densely populated areas, Fig. 3.5 shows that the COMB method produces less accurate results for a large part of the lower population density classes.

3.4 Stability analysis of the parameters

Understanding and quantifying the stability of the estimated parameters α and β is important for the MP method to be applied elsewhere. As outlined in Eq. 3.2, α and β are estimated by using a linear regression on training data to model the relation between MP user density (or MP call density) and population density in each commune. The parameter α represents the ratio between MP user density (or MP call density) and population density, which is adjusted using the census-derived national population. The parameter β reflects the super linear effect of densely populated areas on human activities. Choosing one particular training set over another can thus lead to different estimations of the parameters as different human behaviours or penetration rates can be observed

across regions [190].

Two types of cross-validation procedures are presented here: a standard and a spatially-stratified cross-validation procedure. The range of values obtained for α and β is then used to test the sensitivity of population density estimations to these parameters.

Cross-validation procedures
In the standard cross-validation procedure, 30% of administrative units are randomly sampled and used as a training set to derive α and β coefficients. Accuracy assessment statistics (correlation r and RMSE) are calculated on the independent evaluation set consisting of the remaining 70% of administrative units. The sampling is repeated $1,000$ times in order to provide an assessment of the variability of parameters and accuracy statistics.

Because training and evaluation records are selected at random from the dataset, and population densities are spatially correlated, even a model with poor extrapolation ability may appear to predict well when measured in this way. The ability of a model to make accurate extrapolated predictions in new locations would be better measured by performing a spatially-stratified cross-validation where training and test sets are sampled from geographically distinct regions [191].

We carry out a spatially-stratified cross-validation procedure by assigning administrative units to either the training or evaluation datasets according to whether they fall inside (training) or outside (evaluation) a disc of radius 100 km. Discs are placed at random, centred on the location of an administrative unit, subject to the constraint that the training and evaluation sets contain at least 865 administrative units (30% of the total number of administrative units in Portugal). Below this threshold, the disc radius is iteratively increased or decreased by steps of 10 km until the minimum is reached. This constraint ensures that sufficient data are available to adequately train the model and to evaluate its predictive capacity. The disc-fold validation procedure is implemented in R [192] using code adapted from the sperrorest package [193]. This disc-fold validation procedure is repeated $1,000$ times for each model run, and accuracy assessments are computed (correlation r and RMSE).

Variability of α and β according to the cross-validation procedure

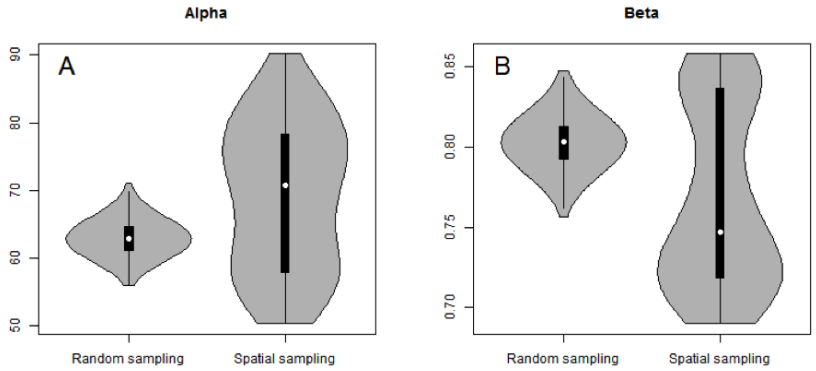


Figure 3.6: **Coefficient variability according to validation procedures. (A)** Alpha and **(B)** beta coefficients estimated using randomly sampled and spatially-stratified training datasets

A best-fit estimate of 62.95 ± 2.48 is found for the parameter α when using a random cross-validation procedure, while this estimate becomes 69.11 ± 10.49 when using a spatially-stratified cross-validation procedure (Fig. 3.6A). The parameter β, which captures the super linear effect that may exist between population density and MP user density, is estimated to 0.803 ± 0.015 when using a standard cross-validation procedure and 0.767 ± 0.055 when using a spatially-stratified cross-validation procedure (Fig. 3.6B).

While the random sampling used in the standard cross-validation procedure has the advantages of removing any cultural or economic bias existing between different geographical regions and limiting spatial autocorrelation problems in the data, the spatially-stratified cross-validation procedure enables reproduction of the initial conditions typically faced by a population distribution modeller when applying a model to a data-scarce country where detailed population data are only available for one region and the model therefore needs to be extrapolated to a geographically different region. In terms of accuracy of population density estimations, our analysis shows that the choice of a particular geographical region over another as training data may induce larger variations in global RMSE (686 ± 173) than the use of a random sample of data for training (574 ± 42) (Fig. 3.7B). Differences in correlation coefficient variations between standard and spatially-stratified cross-validation procedures are less significant, with values of 0.873 ± 0.011 and 0.885 ± 0.011, respectively (Fig. 3.7A).

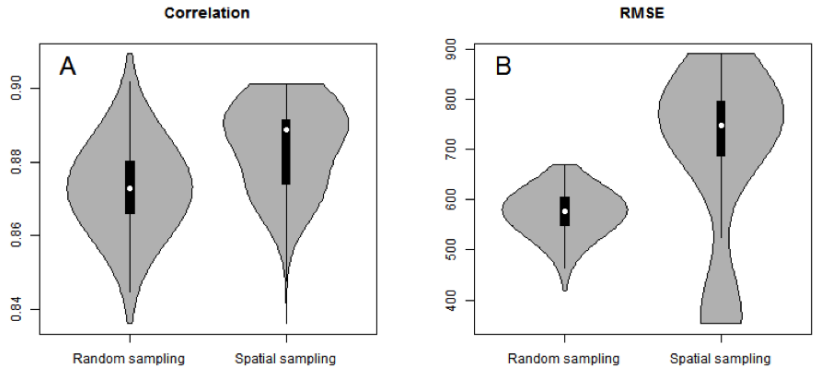


Figure 3.7: **Accuracy according to validation procedures. (A)** Correlation coefficients and **(B)** RMSEs calculated using randomly sampled and spatially-stratified evaluation datasets

When detailed training data exist for calibration, errors can be reduced by choosing a training dataset (i) representative of the larger area to be mapped and (ii) representing a large diversity of population densities. In addition, when allowed by the data, calculating different β coefficients for different regions or different population subgroups should be considered.

Sensitivity of population estimates with respect to α and β
Now that we have a better idea of how α and β values vary according to the training dataset used (Fig. 3.6), it is important to test the sensitivity of population density estimates with respect to the value of these parameters. While the variability of α might seem important, its impact on population density estimations is null, since this parameter is corrected automatically to match the total population of the country (Eq. 3.3). This is confirmed in Figs. 3.8A and 3.8C : when artificially changing the value of α (within the maximum range identified in Fig. 3.6: 50-90), both the RMSE and the correlation coefficient r remain constant.

Unlike α, the sensitivity analysis shows a clear influence of β on the RMSE and r (Figs. 3.8B and 3.8D). A low value of the parameter β means that a proportionally lower population density is assigned to high-density areas compared to low-density areas, which can create large discrepancies in population density estimations, with overestimated population densities in rural areas and underestimated population densities

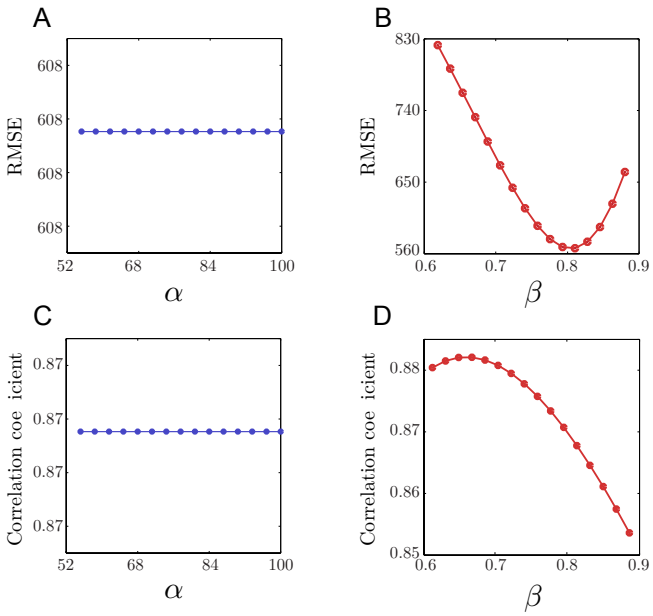


Figure 3.8: **Influence of coefficients on accuracy.** Influence of α and β parameters on the global RMSE and correlation coefficients

in urban areas. A large value of β results in the opposite effect: overestimation of densely populated areas and underestimation of low populated areas, resulting in an increasing global RMSE. In Figs. 3.8B and 3.8D, β values range between 0.69 and 0.86 (maximum range identified in previous section). When using values of β within the confidence interval of 0.77 ± 0.055 obtained with the spatially-stratified cross-validation procedure described above, RMSE values range between 565 and 655 (15% increase) and r ranges between 0.88 and 0.854.

3.5 Flexibility and extrapolation capacity

So far, we estimated population densities from night-time MP user densities. However, other types of input data derived from MP calls can be used as well. In this section, we present a collection of analyses performed to test the flexibility of the MP method in terms of input data used, the impact of potential socio-economical bias and the extrapolation capacity of the method to other countries. First, we test the ability of the density of phone towers, the density of daily-aggregated data and

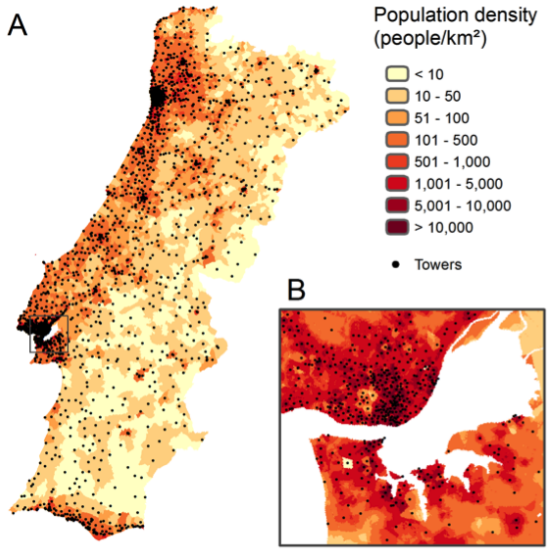


Figure 3.9: **Distribution of MP towers.** Spatial distribution of MP towers **(A)** in Portugal, with **(B)** close-up around the capital city Lisbon. Census-derived population densities are shown in background.

the density of MP calls to accurately estimate population densities. Second, we measure the spatio-temporal variability in phone usage among the mobile phone users and discuss its potential bias on the MP method. Finally, we discuss the extrapolation capacity of the method by applying it to a different country (France).

Density of MP towers

The density of MP towers by administrative unit, denoted by t_{c_i}, is computed with the following equation:

$$t_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} t_{v_j} A_{c_i \cap v_j} \tag{3.5}$$

where A_{c_i} is the area of administrative unit c_i, $A_{c_i \cup v_i}$ is the intersection area of commune c_i and the Voronoi polygon v_j, and t_{v_j} is the MP tower density of Voronoi cell v_j, i.e. $1/A_{v_j}$. In Portugal, the density of MP towers is highly correlated to census-derived population densities ($r = 0.794$), which suggests that using only the density of MP towers would already provide reasonable population density estimates (Fig. 3.9).

Here we compare population mapping accuracies when using the MP method, but using the density of MP towers instead of the density of

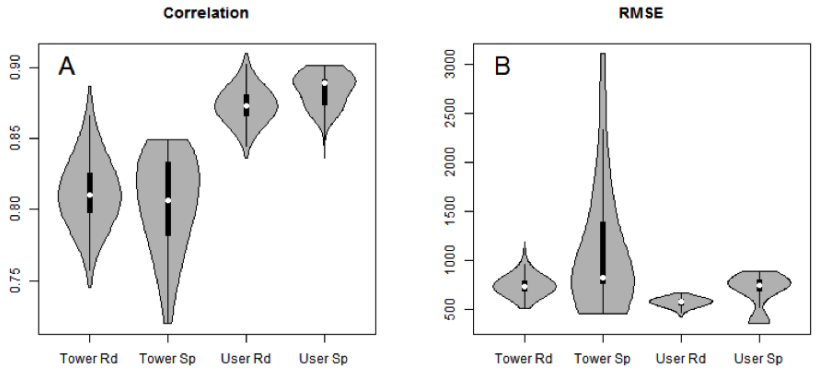


Figure 3.10: **Mapping accuracy associated to MP tower density. (A)** Correlation coefficients and **(B)** RMSEs calculated using the density of phone towers and the density of users (Rd = standard cross-validation procedure; Sp = spatially-stratified cross-validation procedure).

nighttime MP users as input data. Results show that population density estimates are significantly less accurate when using the density of MP towers (Fig. 3.10), with maximum RMSE values being particularly high ($> 3,100$) when using a spatially-stratified cross-validation procedure. In addition, the use of MP towers alone does not allow any dynamic mapping over time.

Daily aggregated data and density of MP calls

The accuracy of the RS method presented in section 3.3 was illustrated with the density of different MP users during the night (8 p.m. to 7 a.m.) as input data. However, network providers do not always provide users' identifiers and the time of phone calls as such details might endanger users privacy. We therefore compute the precision and accuracy of (i) population density estimates obtained from daily-aggregated data compared to nighttime data and (ii) population density estimates derived from MP calls compared to MP user counts. The goal is to evaluate the ability of very basic and fully anonymized MP data to predict human population densities (Figs. 3.11 and 3.12).

Statistical analyses including the analysis of variance and Tukey's honest significant difference tests are performed to test for differences between the different datasets used as input data (see Appendix A for more details on these tests).

Even if the density of calls and the density of users are very highly correlated in Portugal ($r = 0.99$), results show that population density estimates produced using the density of users are generally more precise and accurate than estimates produced from the density of calls (Fig. 3.11CD). However, non-significant differences in RMSE are observed between nighttime calls (CALL NIGHT) and nighttime users (USER NIGHT) ($q_t = 2.62$; $p = 0.24$) (Fig. 3.11D), suggesting that, during the night, using the density of calls instead of the density of users does not impact significantly the accuracy of population density estimates and that the number of calls per user is relatively stable during the night. Note that the we reach the same conclusion when using a spatially-stratified procedure (Fig. 3.12CD).

Results also show that population density estimates produced using nighttime data were significantly more precise and accurate than estimates produced from daily-aggregated data, with r and RMSE statistics being significantly different ($p<0.001$, Fig. 3.11CD). However, the accuracy assessment was done using census-derived nighttime data as a reference, which is not entirely appropriate. For a more precise accuracy assessment, we would need daytime census data as a reference. Nevertheless, the estimated β values between both day/night and call/user data are very close (Figs. 3.11B), which suggests a minimal impact on predicted population densities. When available MP data only include the daily-aggregated number of phone calls (without information on the number of users or on the calling time), as is the case in France, the daily-aggregated number of phone calls can reasonably replace the number of users per night, as long as phone usage behaviors are relatively stable across space and time. The spatio-temporal variability in phone usage is assessed hereafter for Portugal.

Spatio-temporal variability in phone usage

In order to assess the variability of phone usage behavior in time and space, MP users are divided into three distinct profiles, each containing about a third of the total number of users (Fig. 3.13). The profiles are based on the number of phone calls they performed at night during the studied period of 242 days: (i) type 1 corresponding to low-activity users with less than 13 calls (0.054 per night), (ii) type 2 corresponding to medium-activity users with number of calls between 13 and 68 ($[0.054, 0.28]$ per night), (iii) type 3 corresponding to high-activity users with more than 68 calls (0.28 per night). We then analyse how the proportion of users of type 1, type 2 and type 3 vary both in time (Figs. 3.14 and 3.15) and in space (Figs. 3.16, 3.17 and 3.18).

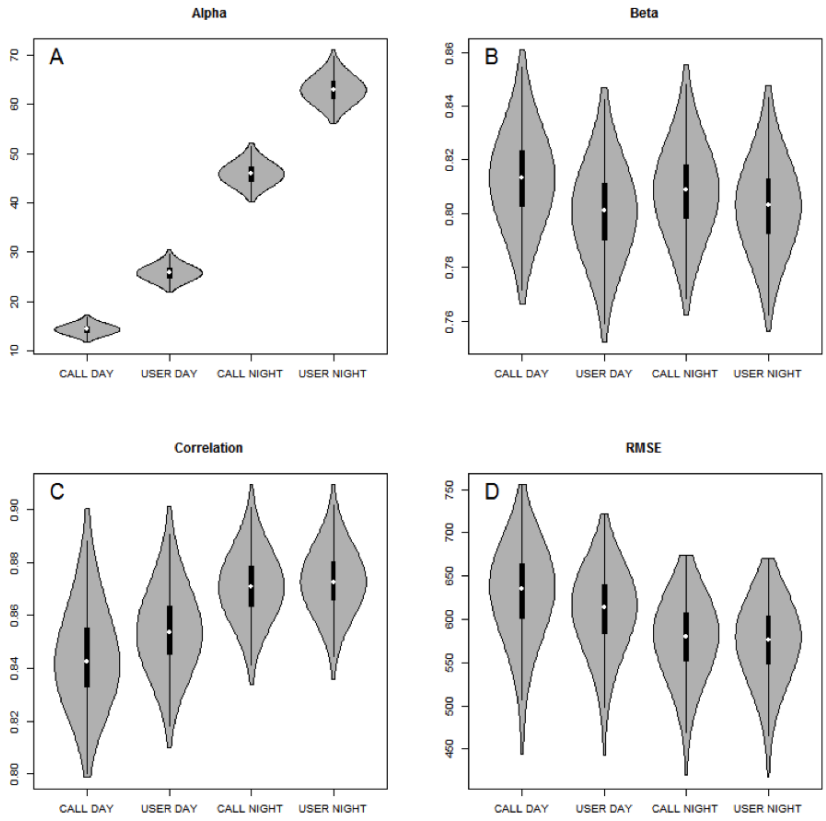


Figure 3.11: **Variability and accuracy associated to different input data for standard cross validation procedure. (A)** Alpha, **(B)** beta, **(C)** correlation coefficient and **(D)** RMSE calculated when using (i) daily-aggregated calls (CALL DAY), (ii) daily-aggregated users (USER DAY), (iii) nighttime calls (CALL NIGHT) and (iv) nighttime users (USER NIGHT), with a standard cross-validation procedure.

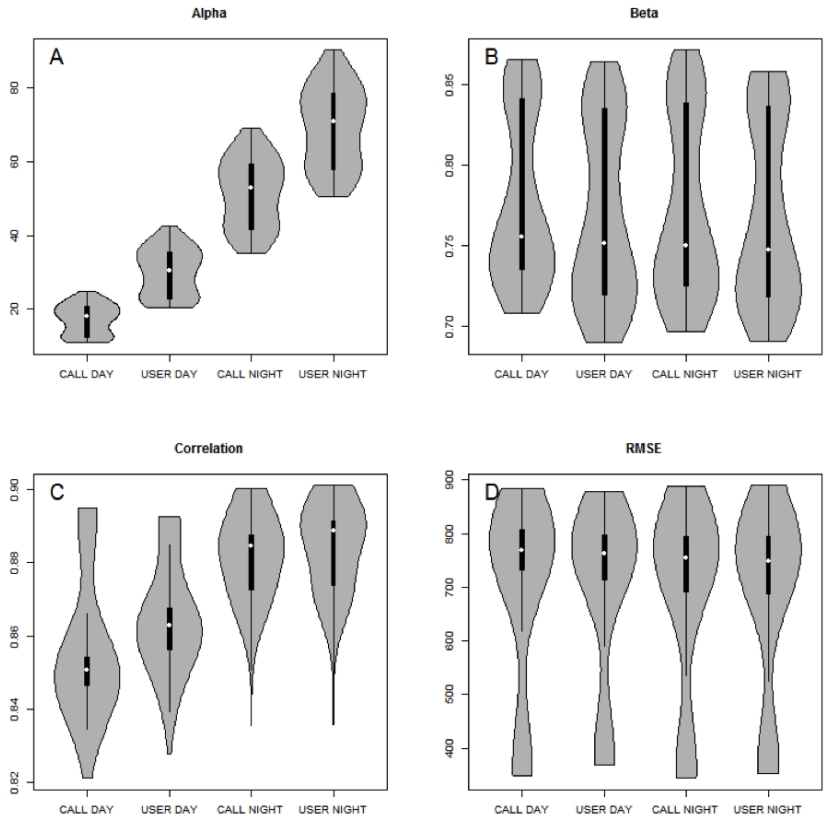


Figure 3.12: **Variability and accuracy associated to different input data for spatially-stratified cross validation procedure. (A)** Alpha, **(B)** beta, **(C)** correlation coefficient and **(D)** RMSE calculated when using (i) daily-aggregated calls (CALL DAY), (ii) daily-aggregated users (USER DAY), (iii) nighttime calls (CALL NIGHT) and (iv) nighttime users (USER NIGHT), with a spatially-stratified cross-validation procedure.

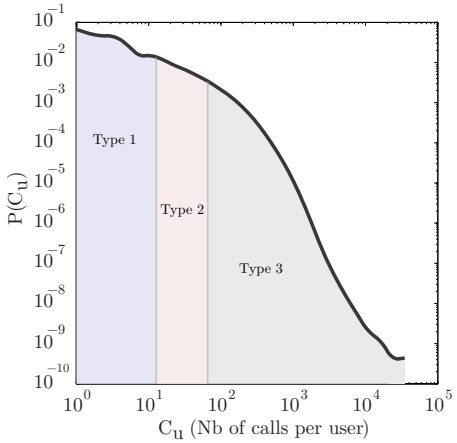


Figure 3.13: **User profiles.** Probability Density Function of total number of night phone calls per user. Mobile phone users are divided into three distinct profiles, each containing a third of the users: low-activity users (Type 1), medium-activity users (Type 2) and high-activity users (Type 3).

Results show that the proportion of each profile is stable over the week (Fig. 3.14), but less over the day (Fig. 3.15). Indeed, we observe that the proportion of high-activity users (type 3) is lower during the day than during the night while the proportion of low and medium-activity users (types 1 and 2) is higher during the day than the night. Considering day-time and night-time data separately, as we do in our manuscript, is thus important in order to study users with stable behaviors.

To analyze the variability in the proportion of users of type 1, type 2 and type 3 in space, we used three variables that are spatially clustered: the population density (Fig. 3.16), the unemployement rate (Fig. 3.17) and the percentage of people who hold a higher education degree (Fig. 3.18). These data were obtained from the National Institute of Statistics of Portugal by administrative unit level 5 (ADM-5) for the year 2011 [185] and were aggregated to Voronoi polygons corresponding to phone towers. Fig. 3.16 shows that the proportion of each user profile varies across space, with a higher proportion of high activity users (Type 3) than low and medium activity users (Type 1 and 2) in densely populated areas. This well-known super-linear effect of population density on human activities is however captured by the coefficient β in our model.

The proportion of each user profile also varies with the proportion

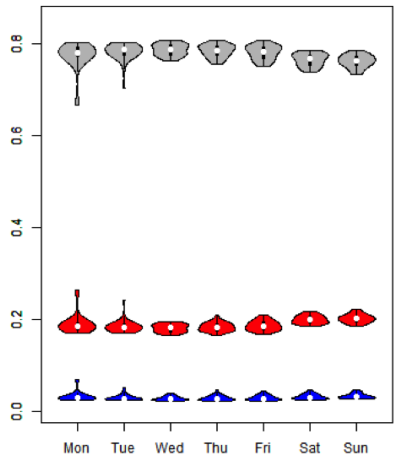


Figure 3.14: **Variability of user profiles over time (day).** Distribution of proportion of users of type 1 (blue), type 2 (red) and type 3 (grey) for each day of the week.

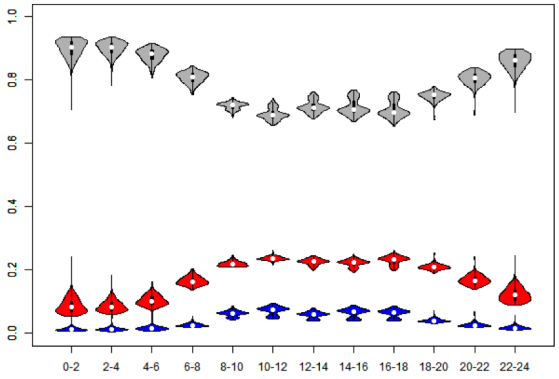


Figure 3.15: **Variability of user profiles over time (2 hours).** Distribution of proportion of users of type 1 (blue), type 2 (red) and type 3 (grey) for each 2-hour period of the day.

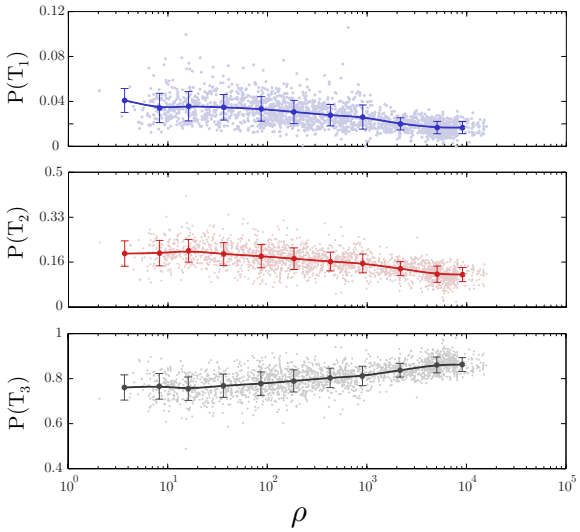


Figure 3.16: **Variability of user profiles according to population density.** The proportion of low (blue) and medium (red) activity users (type 1 and 2) tend to decrease in densely populated areas, while the proportion of high-activity users (grey) increases (type 3).

of people holding a higher education degree (Fig. 3.17), with a larger proportion of high activity users (Type 3) in administrative units where the proportion of people holding a higher education degree is higher. However, this trend is mainly due to the correlation between the population density and the higher education degree ($r = 0.52$), which suggests that the influence of the education level is captured by the coefficient β. There is however no clear relation between the proportion of each user profile and the unemployment rate at the mobile phone tower level (Fig. 3.18), suggesting that unemployment rate does not influence the mobile phone behavior of users in Portugal.

Stability with respect to other countries

The population downscaling method developed in this chapter was applied to France. Instead of the number of different users per night, we used here the number of daily-aggregated calls made or received from each tower during working periods (May, June, September, October 2007) to train the model. We have seen in the previous subsections that using daily-aggregated call data had an impact on accuracy statistics, though this impact was largely due to the use of residential census data as reference for the accuracy assessment. However, this impact was rather low and not always significant.

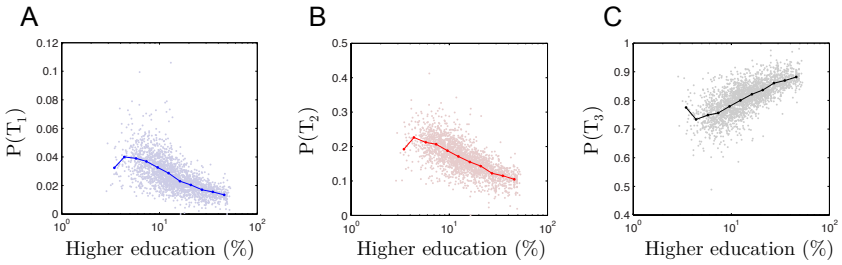


Figure 3.17: **Variability of user profiles according to higher education.** Variability of user profiles at each mobile phone tower according to the percentage of people holding a higher education degree. The proportion of **(A)** low and **(B)** medium activity users ($T_1$and T_2) tend to decrease with the education level, while the proportion of **(C)** high-activity users increases (T_3)

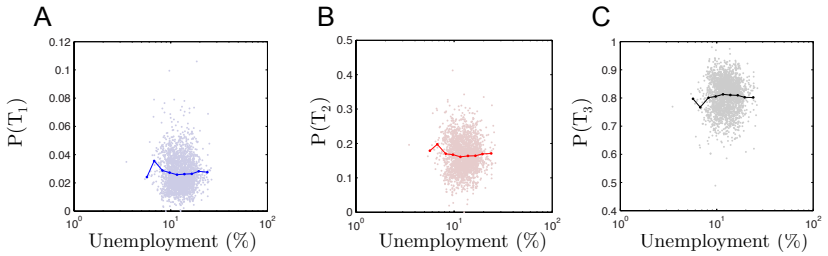


Figure 3.18: **Variability of user profiles according to unemployment rate.** We observe no correlation between unemployment rate and the proportion of **(A)** low, **(B)** medium and **(C)** high activity users.

We compare β coefficients resulting from the French dataset with the values we had for Portugal (using daily aggregated MP call data) in order to assess the variability of this coefficient between countries (Fig. 3.19). The standard and spatially-stratified cross-validation procedures defined in section 6.4. are used to derive α and β coefficients for France. In order to use training datasets of comparable size for Portugal and France, only 2.5% of the $36,610$ administrative units available for France are used as training data. Results show that β is higher in France (0.902 ± 0.036) than Portugal (0.813 ± 0.016) when estimated using a standard cross-validation procedure (Fig. 3.19A), but confidence intervals largely overlap when they are estimated using a spatially-stratified cross-validation procedure, with β values of 0.777 ± 0.051 for Portugal and 0.846 ± 0.056 for France (Fig. 3.19B). The larger confidence intervals observed for France are due to the higher number of administrative units available and the resulting greater diversity of administrative units sampled for training models.

In France, two regions (Corse and Provence-Alpes-Cote-d'Azur) are characterized by a particularly high proportion of tourists, with rates of camping area per person being the highest for these two regions (0.07 and 0.02 for the region of Corse and Provence-Alpes-Cote-d'Azur respectively, while the national average is 0.01) [186]. When using these regions as training datasets, estimated β values are above 1, suggesting that a higher proportion of calls are made in low-density areas than in high-density areas in these regions. If we exclude these two regions from the training datasets, estimated β values are slightly lower (0.894 ± 0.035 with a standard cross-validation procedure and 0.842 ± 0.046 with a spatially-stratified cross-validation procedure). Choosing a training dataset that excludes the main holiday periods and typical tourism areas should thus be considered to reduce errors in population density estimates. It would indeed limit the discrepancies between residential and temporary population distributions.

3.6 Population dynamics

One of the main advantages of the MP method is its ability to be applied on different time periods and thus to capture population dynamical changes in a region or a country over time. Temporal dynamics of population densities can be derived from MP data using the timestamp associated to each MP call. Three distinct temporal dynamics are

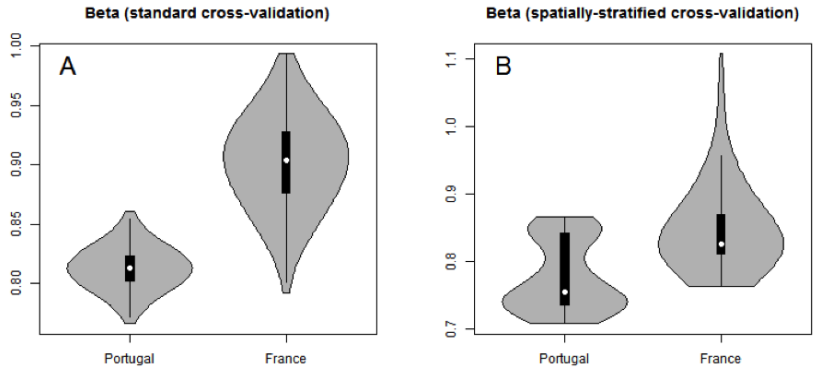


Figure 3.19: **Variability of coefficients for France and Portugal.** Comparison of β estimations in Portugal and France using **(A)** a standard cross-validation procedure and **(B)** a spatially-stratified cross-validation procedure.

considered here based on MP calls densities: (i) daily dynamics are analysed by dividing the MP data into MP calls performed during the day (7 a.m. to 8 p.m.) and the night (8 p.m. to 7 a.m.), (ii) weekly dynamics are obtained by dividing the MP data into MP calls performed during weekdays (Monday to Friday) and MP calls performed during weekends (Saturday and Sunday) and finally (iii) seasonal dynamics are obtained by dividing MP data into MP calls performed during the holiday period (July and August) and MP calls performed during working periods (all other months). For each temporal dynamic, predicted population densities for each unit and for both time periods can be computed using best-fit α and β estimates. The relative differences of population densities within each area and between the two time periods can then be computed to detect the fluctuations.

The potential of MP data to estimate population density variations through time is illustrated in Fig. 3.20 for the seasonal dynamics. The relative differences in estimated population densities between the major holiday period (July and August) and more traditional working periods (from September to June) in Portugal and France reveal clear spatial patterns (Fig. 3.20). Seasonal changes in population distribution are evident: most cities are characterized by a large decrease of population densities during the holiday period, while less populated areas and well-known tourist sites such as coastlines or mountainous areas show large increases. Fig. 3.20E shows that population densities decrease in Paris,

with the exception of a few spots corresponding to highly visited sites (e.g. Disneyland Paris, Charles de Gaulle airport). A comparison of the population density variations for the other temporal dynamics is illustrated as well in Fig. 3.21 for Portugal. Results show again clear spatial patterns: population density increases along highways during the day (Fig. 3.21A), population density decreases in major cities during both weekends and holidays (Figs. 3.21B,C) and population density increases along the coast during holidays (Figs. 3.21C).

The observed differences may be influenced by the variations in phone usage behaviors mentioned in the previous section. During the day, the proportion of low and medium-activity users is higher in (Fig. 3.15), resulting in a lower number of phone calls per user. Such day/night variations may be therefore more visible when using the number of users than the number of calls. This emphasizes that, when data include users' identifiers, it is preferable to use the number of users than the number of calls. Some other phone usage behaviors may influence day/night variations such as the use of professional phones during the day and private phones during the night. This suggests that estimates may become more uncertain over shorter timescales.

Finally, a positive correlation is observed between the difference in estimated population between the holiday and the working periods and the number of tourist accommodations available by commune ($r = 0.28$) [186]. In addition to providing quantitative measures of how people from densely populated areas tend to travel towards low-density and recreational locations during holidays or weekends, this method also offers a detailed visualisation and quantification of the dynamic popularity of a given place over time.

3.7 Conclusion

The increasing penetration of mobile phones and other information and communication tools used daily by a high proportion of the global population offers a wealth of new spatio-temporal data that is contributing to the *big data* revolution. These new data have the potential to profoundly transform the way we think about and conduct science, especially geographical analyses, as most of these data are implicitly or explicitly spatial [194, 195]. In operational and governmental decisions these data may also be valuable for supporting rapid responses to disruptive events

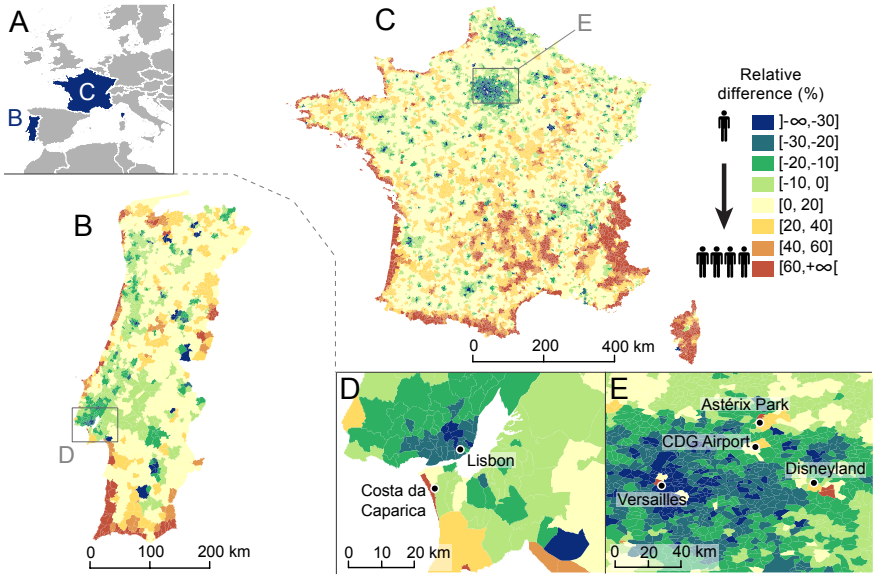


Figure 3.20: **Population dynamics in France and Portugal (A)** Location of Portugal and France in western Europe. **(B-E)** Relative difference in predicted population density between the main holiday period (July and August) and the working period (September-June) by administrative unit level 5 (ADM-5) **(B)** in continental Portugal, **(C)** in metropolitan France, with **(D)** a close-up around Lisbon, with labels showing the city center of Lisbon and the seaside resort "Costa da Caparica" and **(E)** a close-up around Paris, with labels showing the busiest airport in the country (Paris Charles de Gaulle), one of the most visited places in France (Palace of Versailles) and two popular recreation areas (Disneyland and Asterix Park)

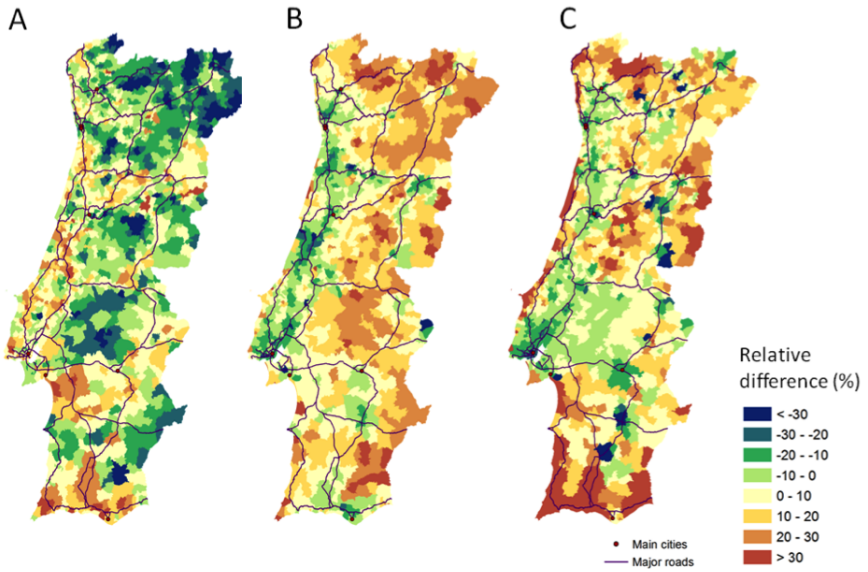


Figure 3.21: **Population dynamics in Portugal for different time windows.** Relative difference in predicted population density by ADM-5 for different time periods in Portugal. **(A)** Difference between day and night, with brown colors indicating a higher population density during the day; **(B)** difference between weekend and weekdays, with brown colors indicating a higher population density during weekends; **(C)** difference between the main holiday period (July and August) and the working period (November-May), with brown colors indicating a higher population density during the holidays.

or longer term planning purposes. In the specific application presented in this chapter, spatially and temporally detailed population distribution datasets can potentially provide the essential denominator required in many fields, such as studying collective human responses to disease outbreaks [196, 197], emergencies [198, 199] or any application where information on daily, seasonal or annual changes in population distribution are of use.

This chapter demonstrates how the analysis of MP data that is readily collected every day by phone network providers can complement traditional census outputs. Not only can population maps of comparable accuracy to census data and existing downscaling methods be constructed solely from MP data, but these data offer additional benefits in terms of the measurements of population dynamics. Further, as highlighted in section 3.3, a combination of both the MP and RS methods facilitates the improvement of both spatial and temporal resolutions and demonstrates how high resolution population datasets can be produced for any time period.

In countries where detailed human population census data are available at high resolution, the main value added is not so much in the gain in spatial resolution, but more in the ability to estimate population numbers and densities at high spatial resolution for any time period. This allows us to follow how population distribution changes through time in relation to the week, the season, or any particular event affecting populations over large spatial extents. The relevance of the MP approach is even higher in low-income countries where population distribution data can be scarce, outdated and unreliable. In Africa, great variation exists in the quality of spatially-referenced population data. In Malawi for example, censuses have been performed once per decade for the last three decades and data are readily available at the level of enumeration areas (i.e. administrative units of 9.38 km^2 on average). In contrast, in the Democratic Republic of the Congo (DRC) the most recent census was undertaken in 1984 and data are only available at the level of Territories (i.e. administrative units of 12,466 km^2 on average). However, in the DRC the MP penetration rate, though biased towards certain demographic groups, is relatively high (69% on average by the end of 2014 in Africa [175]) and the MP approach would produce considerable improvements to current knowledge of how population is distributed in the country. Even if at present the most remote and isolated populations may not have reception in some low income countries, and thus may affect the ability to produce a comprehensive country-wide map,

network coverage continues to grow at a rapid rate everywhere.

Applying the approach to countries such as the DRC, where reliable training data may not be available, requires some adjustments and assumptions, in particular regarding the relation between the MP user density and the population density, through estimates for the parameters α and β. This relation can indeed vary between and within countries according to the penetration rate of the network operator and phone usage behaviors. Network access costs and cultural differences between countries can, for instance, result in communication via text messages being preferred over calls in some countries. Such differential phone usage between countries could largely be accounted for by adjusting total populations using national population counts. A further complication is that phone usage and penetration rates are rarely uniform within countries. In France, the general penetration rate varies from 62.8 in the Franche-Comté region to 117.9 in Ile-de-France, according to the ARCEP [190]. Such regional MP ownership information are generally available either from independent bodies such as regulators, phone operators themselves, or can be estimated through national household surveys such as the Demographic and Health Surveys [200] and give a first indication of potential phone usage variations between regions. The spatially-stratified cross-validation procedure used here enables assessment of the impact of regional variations on model parameters in Portugal and France, and the impact of such variation on population mapping accuracies (section 3.4). Spatial variations in phone usage behaviors can also be due to economic, social, demographic or cultural characteristics that can be spatially-clustered and therefore bias population density estimates. While a complete analysis of such potential biases is beyond the scope of this chapter, we showed that phone usage behaviors were relatively stable across space and time in Portugal and that a large part of the variation is correlated to population density, and is therefore captured by the coefficient β (section 3.5).

In order to be widely applied and to facilitate the acquisition of MP data, the method outlined here can be simplified by using the density of phone calls instead of the density of different users over a certain time window. Even if resulting population density datasets are marginally less accurate, this allows the method to become independent from user identifier data and further reduce privacy concerns (section 3.5). Similarly, using daily-aggregated data instead of night data again reduces marginally the accuracy of estimates, though notably simplifying the acquisition and processing of MP data.

The observed robustness of the MP method offers promise for extension of the mapping to other countries and network providers. However, applying the method to low-income countries where penetration rates are rapidly increasing but still exclude an important fraction of the population would require further sensitivity analyses of the impact of phone usage inequalities, especially as marginalized populations are also the most vulnerable to disasters, outbreaks or conflicts. Mobility estimates in Kenya were found to be surprisingly robust to the substantial biases in phone ownership across different geographical and socioeconomic groups [201], but these results would need to be confirmed for population density estimates.

Mobile phone call data records are constantly collected by network providers, but the potential of such data are only demonstrated sporadically. A wider use of such data is currently impeded principally by privacy and data access concerns. The use of call data records does raise important privacy concerns linked to fundamental questions of personal freedom and ethics. Studies of individual mobility patterns provide little anonymity, as the movements of individuals can be reconstructed in time and space, even if spatially and temporally coarsened datasets are used [202]. Here by using only phone call activity aggregated by tower, neither individual data nor connections between towers are used, guaranteeing the privacy of MP users. A facilitated access to anonymized and aggregated forms of these data would greatly improve our knowledge of human population distributions and movements. Network providers are sometimes reticent to share their data because of privacy and marketing concerns. However, this study has shown that aggregated and anonymized MP data could cost-effectively provide accurate maps of population distribution for every country in the world for every month. Partnerships between governments and phone companies supported by appropriate incentives could enable fast and cheap production of population maps in emergency contexts, enabling rapid assessments of populations at risk, or impacted by disasters, disease outbreaks or conflict.

Chapter 4

Spatial distribution of social communities

In this chapter, we investigate the spatial distribution of communities in social networks as well as their stability over space and time. First, we describe two geographical social networks built from mobile phone data. Second, we introduce a community detection method that maps communities in these two spatial networks. Third, we present the resulting community partitions and discuss their stability over space and time. Finally, we present a novel technique that aims at quantifying the spatial stability of a particular node within a community and discuss its potential applications.

4.1 Introduction

Social, technological and information systems can often be described in terms of complex networks [51, 203–205]. One of the most useful analyses one can do on a such large networks is community detection which consists of decomposing the networks into almost connected subnetworks, called communities [206, 207].

Communities emerge in many networked systems in social sciences [208, 209], biology [210–213], ecology [214, 215], computer science [216, 217], but also economics [218] and politics [205]. They allow a better understanding of these networks because communities often have a clear meaning. In online social networks communities often identify clear virtual groups of users sharing the same interest [208, 209], in the network of the World Wide Web, they group web pages related to the same topic [217], in protein-protein networks, they are likely to put together proteins having a similar function within the cell [211–213] or

in food webs they may correspond to compartments, i.e. subgroups of taxa [215]. Communities also fuel numerous applications, from improving performance of services on the Web by grouping clients that are geographically closer to each other and share similar interests [219] to interpreting genome-wide association study by identifying connected subgraphs of proteins associated with the same disease [220].

As many complex systems are organized in the form of a network embedded in space, the detection of communities is also important to uncover their spatial structure. Important examples include social networks, the physical Internet infrastucture, road networks, flight connections and brain functional networks. In social networks, a central issue is to understand the role played by regional boundaries defined by governments on the way people interact across space [27, 221, 222]. This understanding is fundamental in economic geography [36–38], but is also the underlying cause of many social, political and ethnic conflicts across the world [26, 223–225]. Traditional methods have focused on aggregated census, market or travel data to estimate the interactions between and within regions [226–228]. However, this is rapidly changing as new sources of large-scale finer-grained social data such as mobile phone data has become more and more available.

Mobile phone communications provide fertile territory for research into the spatial dimensions of communities. Studies of calling patterns have shed new light on the complex nature of networks. The analysis of billions of calls across a number of countries has led to surprising conclusions such as the existence of geographically cohesive regions that correspond remarkably well with administrative regions, while unveiling unexpected spatial structures [26, 27, 229].

In this chapter, we present the first approach that delineates regions of France based on a dataset of more than a billion phone records. We first introduce two spatial social networks built from phone communications. We then introduce the method used to detect communities in our two networks as well as their resulting partitions into communities. Next, we assess the stability of these communities over time, exhibiting the influence of particular time periods on the resulting partitions as well as the influence of administrative boundaries. We then introduce a sensitivity measure that aims at quantifying the stability of particular nodes within a community and we discuss its potential applications.

4.2 Social network construction

In this section, we describe the process with which we extract information from mobile phone data in order to build two social networks embedded in space. One is based on mobile phone tower locations while the other is based on zip codes of billing addresses. We first describe the preprocessing applied to the data and then present the different aggregations made in order to build the networks.

Preprocessing of the Data
The raw data consists of mobile phone records from the largest operator in France. The data is thus the most representative one could obtain from a single provider. More than 42% of the population (25 million people) are present in the data which cover a period of 5 months from May 2007 to October 2007.

For each phone call, the following information is available

u_i	Z_i	T_i	u_j	Z_j	T_j	t	Δt

where u_i, Z_i and T_i correspond to the user identifier, the zip code of the address and the tower routing the call of the caller and conversely where u_j, Z_j and T_j correspond to the user identifier, the zip code of the address and the tower routing the call of the callee. The time when the call was initiated (t) as well as the duration of the call (Δt) are also known.

In order to obtain accurate information, only data fulfilling several requirements are extracted. The first requirement concerns the availability of customers billing addresses. In order to preserve this information, we only keep calls between users who subscribed to a contract with the operator. Second, we exclude customers who subscribed to pre-paid or professional contracts as the billing address is either not available or corresponds to the address of a company. Finally, only voice calls are taken into account in order to have a single type of communication. After this preprocessing, the data amount to about 1.2 billion phone calls and 17 million users, which is about 27% of the total population of the country.

Social network based on mobile phone tower
In this network the real location of users when they perform a call is

incorporated. When one makes a phone call, the network usually identifies nearby towers and connects with the closest one [187]. As a result, we aggregate the network of users — connected through their phone calls — to a network of towers, where each tower regroups all calls that were passed by users within its vicinity. This tower-based positioning method is simple to implement and its accuracy directly depends upon the network structure; the higher the density of towers, the higher the precision of the MP communication geo-localization [184]. Given that the distribution of mobile phone towers is directly correlated to the population density, higher precision is achieved in densely populated areas such as cities. The resulting network is composed of $17,500$ nodes (mobile phone towers) connected through 1.2 billion phone calls.

Social network based on billing addresses
In the second network, we aggregate the network of users to a network of communes where each commune regroups phone calls passed by users living in that commune. As a result, this network does not consider the current location of users. Indeed, a same user performing multiple calls from distinct locations will appear as static in the network structure as the same location (his home zip code) will be used as origin for all of his calls. Since the spatial distribution of zip codes is more evenly distributed over the country than cell phone towers, the accuracy of this positioning method is stable over space, offering a better accuracy in low populated area compared to the tower based network. This geotagging technique has the advantage of characterising social ties between locations where people actually live in, rather than where they move. This offers a different but complementary picture of the spatial structure of the social network in France. The resulting network is made of $6,000$ nodes (zip codes) connected through a total of 1.2 billion phone calls.

4.3 Community detection

The detection of communities in a network is a very popular but not yet a satisfactorily solved problem, despite the great effort of a large interdisciplinary community of scientists working on it over the past few years. Indeed, the task of detecting communities is very hard, both conceptually and practically. Conceptually because the definition of communities is not well defined and commonly referred to a group of nodes that are strongly connected to each others but weakly connected to other nodes of the network. Practically because algorithms have to find an optimal

partition among an exponentially large number of possibilities. This ambiguity of the definition associated to its complexity gave rise to a rapidly growing scientific community over the past decade as well as myriads of different approaches such as divisive algorithms [206, 230], hierarchical and partitional clustering algorithms [231–233], dynamic algorithms [234–236] or modularity-based methods [229, 237–239]. In this section, we present the algorithm used to derive communities for our two spatial social networks. First we introduce the concept of modularity on which the algorithm relies. We then describe the method, detailing the different optimization steps, as well as discussing its benefits and drawbacks.

Concept of Modularity

Modularity, denoted as Q, is a scalar value that measures the quality of a given node partition of a network. Intuitively, the modularity measures the density of the links inside communities as compared to inter-community links [240, 241]. In this chapter, since we are dealing with weighted networks, this measure is defined as

$$Q = \sum_{i=0}^{n} \left(\frac{l_{C_i}}{l_G} - \left(\frac{d_{C_i}}{2l_G} \right)^2 \right) \tag{4.1}$$

where l_G is the sum of the weights of all links in the network, d_{C_i} is the sum of weights of links incident to nodes in community C_i, and l_{C_i} is the sum of the weights of links inside C_i. More precisely, this quantity measures the fraction of links in the network that connect nodes of the same community minus the expected value of the same quantity in a network with the same partition but where links are randomly shuffled. If the number of edges within communities is similar to what we would get in a random case, we get $Q = 0$. Values approaching $Q = 1$, which is the maximum, corresponds to a network partition with a strong community structure while $Q = -1$, the minimum, corresponds to a partition where communities are non-existent.

Initially introduced by Newman *et al.* [240], this measure was first used to assess the quality of a partition and compare partitions of a network. Later, modularity has also been used as an objective function to optimize [237] and has rapidly become an essential element of many clustering methods [229, 237–239]. However, it has been shown that modularity has a significant drawback — called the *resolution limit* — as the size of the optimal communities depends on the size of the net-

work [242].

The Louvain method

Recently, Blondel *et al* introduced an algorithm of community detection called the *Louvain method* [229]. This method has several advantages such as its simplicity of use and its fast computation time on large networks like the ones we are dealing with. The algorithm of Louvain is a modularity-based method. Its principle is quite intuitive and is divided in two main phases repeated iteratively as follow

Louvain Algorithm

Given a weighted network of N nodes,

Initialization Each node n_i is assigned to a distinct community.

Phase 1 For each node n_i, we consider all neighbors n_j of n_i and we evaluate the gain of modularity resulting from the removal of node n_i from its community and its insertion to the community assigned to one of the node n_j. The node n_i is then assigned to the community for which the gain is positive and maximum. This phase is repeated iteratively until no improvement of the modularity is possible for every individual node.

Phase 2 A new network whose nodes correspond now to the communities found during *Phase 1* is built. The weights of the links between two new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities. The links between nodes within the same community lead to self loops. The *Phase 1* is then repeated on this new network.

The process iterates *Phase 1* and *2* until no more improvement of the modularity is possible after *Phase 2*.

Results : Returns a hierarchical partition of the network.

This algorithm provides a decomposition of the network into communities for different levels of organization and incorporates a notion of hierarchy as communities of communities are built during the process.

This modularity maximisation technique is quite efficient and allows one to analyse very large complex networks [243, 244].

Although widely used in practice, the behavior and accuracy of the technique of modularity maximization is not well understood in practical contexts. A broad characterization of its performance has been made showing that resulting solutions should be interpreted cautiously in scientific context and that this functions admits several local optima [245]. Given these facts, several analyses are performed in this chapter to evaluate in detail the stability and the accuracy of the resulting partitions.

4.4 Spatial distributions of communities

In this section, we first present the resulting partitions of the two social networks introduced in section 4.2 and embedded in space. We then address the issue of stability of the partitions over time and discuss the influence of administrative boundaries on their spatial configurations.

Matching between communities and administrative regions
To facilitate the comprehension of the spatial structure of communities throughout this chapter, we use a detailed caption for the administrative regions of France given in the following table and in Figure 4.1.

1	Nord-Pas-de-Calais	**12**	Bourgogne
2	Picardie	**13**	Franche-Comté
3	Haute-Normandie	**14**	Poitous-Charente
4	Champagne-Ardenne	**15**	Limousin
5	Lorraine	**16**	Auvergne
6	Alsace	**17**	Rhône-Alpes
7	Basse-Normandie	**18**	Aquitaine
8	Ile-de-France	**19**	Midi-Pyrénés
9	Bretagne	**20**	Languedoc-Roussillon
10	Pays de le Loire	**21**	Provence-Alpes-Côte-d'Azur
11	Centre	**22**	Corse

The resulting communities from the Louvain method for both social networks is represented on Figure 4.2. The tower-based and zip-based social networks are characterised by a modularity of $Q = 0.79$

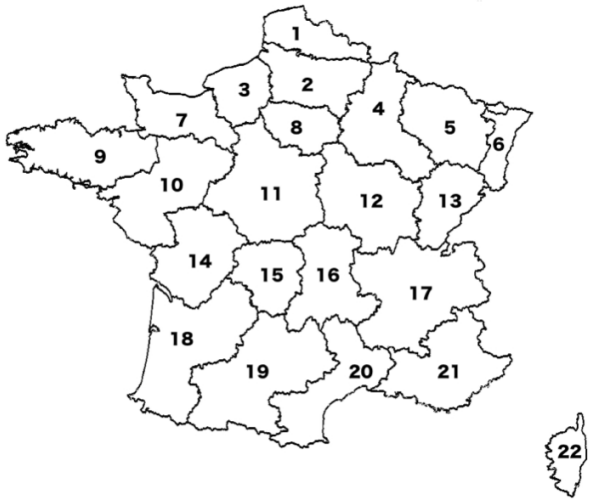


Figure 4.1: **Numbering of Regions of France**

and $Q = 0.83$ respectively. To begin, the uncovered communities adhere to distinct spatial logic although no geographical presupposition is made in the data. Communities align themselves along rather strict borders and their populations are generally uniform (as opposed to being spread across numerous disjointed pockets). Such configuration is not surprising as there has been much empirical evidence about the geographic impact on communication patterns [246, 247], documenting that the probability for two individuals to communicate decays with distance, following power law distributions [248, 249]. Aside from a small number of outliers, communities are divided according to previously defined administrative boundaries with a surprising level of precision. It should also be noted that while most regional boundaries contain a single homogenous community, others contain multiple groups. This should come as little surprise as French regions are a fairly recent invention and as Figure 4.2 clearly demonstrates local populations are more deeply attached to the more ancient notion of their department, thus adding further weight to the argument that communication patterns tend to follow well-established administrative boundaries. It is also important to mention that this surprising spatial community configuration is robust against the method used. Indeed, while we chose the Louvain method in this chapter, a strong influence of administrative units on social communities is also observed when using other community detection methods on the same social network [250], reinforcing our results.

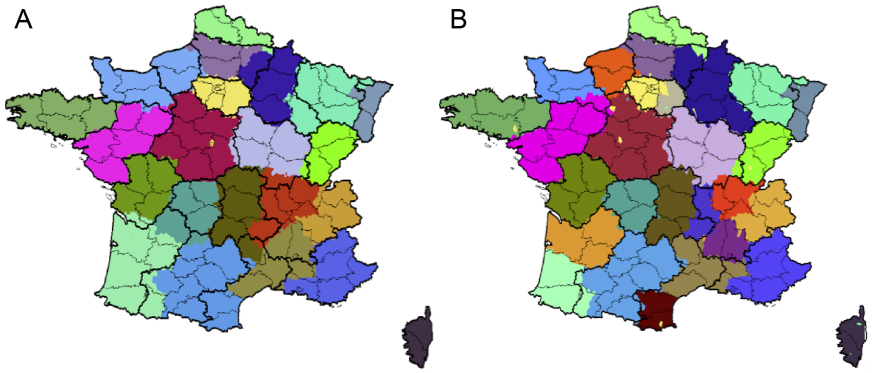


Figure 4.2: **Spatial distribution of communities in France.** Visualization of the communities over a 5 month period for the network based on **(A)** mobile phone tower and **(B)** zip codes. Each community is characterised by a specific colour while administrative regions are delineated by thick black lines.

While the influence of administrative borders on communities in large social networks is known [26, 27] such correlation with administrative regions is still surprising. Regional linguistic variations offer one logical explanation for the divisions observed in Belgium [26] and to a somewhat lesser degree in Britain [27] but in France no such claim can be made. It would be reasonable to expect therefore some differences to the examples of Belgium and Britain but the actual results contradict this assumption. Indeed, if anything, France demonstrates an even starker division along administrative lines.

Spatial stability over time

The obtained communities (Fig. 4.2) offer only a static snapshot of the network over a 5-month period. However, population distributions and social interactions are dynamic over time [59]. It is thus important to consider and evaluate the robustness of these partitions over time as the network is evolving.

To assess the robustness over time of a given partition one could divide the data into different time windows and evaluate the quality and consistency of the detected communities in the resulting time-varying network [251]. Comparing the spatial stability of communities over time can then be assessed by superposing the geographical delimitation of communities obtained for the different time windows. The superposition of the weekly partitions reveals a rather stable picture of communities

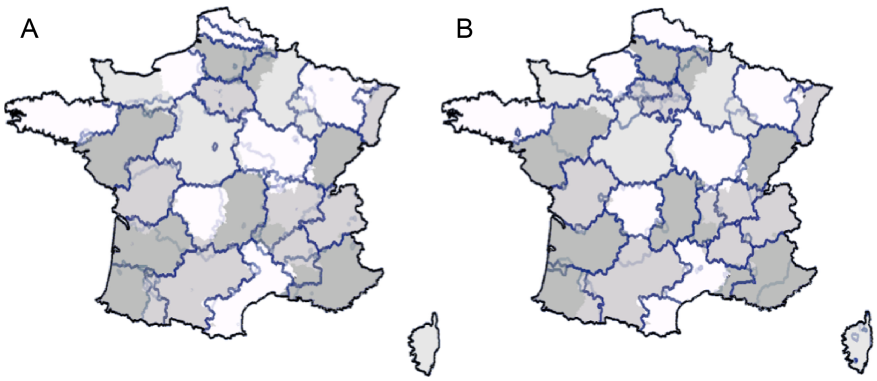


Figure 4.3: **Spatial stability of communities in France.** Superposition of community borders of weekly partitions for **(A)** the tower-based network and **(B)** the zip code-based network. Administrative regions are delineated by grey areas while community partitions are delineated by superposed blue lines.

in France (Fig. 4.3), confirming the clear division along administrative lines where darker colour are present.

Beyond demonstrating the stability between partitions, this type of visualization is also useful to detect particular communities that are likely to be more unstable than others. To understand the source of these slight variations over time, one can compute the modularity values corresponding to weekly partitions (Fig. 4.4). While the modularity values are stable during work periods (May-June and September-October), we observe a significant drop for the holiday period (July-August). However, this phenomena is not surprising as populations tend to change their social and mobility habits during holidays [59, 129, 252]. The structure of both networks is thus changing over time, especially for the tower-based network as it incorporates the mobility of users. In the next section, we present a method that can cost effectively assess the global spatial stability of a particular community as well as the stability of particular nodes.

4.5 Sensitivity measure

As presented in the previous section, some communities seem less stable than others: they merge, divide and sometimes disappear. One of the

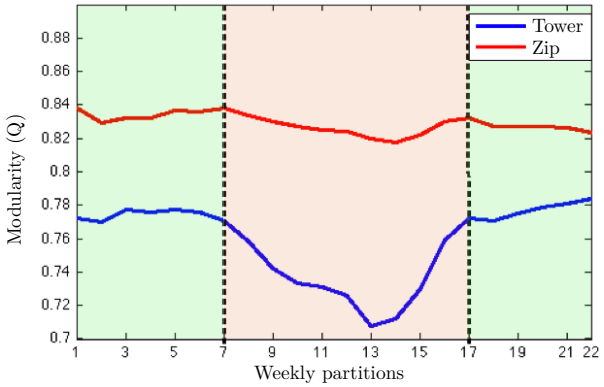


Figure 4.4: **Modularity values of weekly partitions.** Evolution of the modularity values for different weekly partitions for both social networks. A drop of modularity is observed for partitions corresponding to the holiday periods (red zone), while it remains stable for both work periods (green zones).

reasons for this instability is because nodes often belong to more than one community, resulting in overlapping communities [253–257]. This instability can also result from the definition of the modularity function as it has been showed that it may admit several local optima [245]. To take into account these phenomena, we introduce a *sensitivity* measure that evaluates the connection strength between a node and its assigned community. Beyond its ability to slightly improve the quality of a given partition in a network in terms of modularity, this measure can assess the spatial stability of community borders, highlighting the spatial fracture of social interactions between contiguous administrative regions in France.

Method

The method used to build this sensitivity measure is based on the modularity function (Eq. 4.1). The main idea is to observe the fluctuation of the modularity function when moving a node from its community to a neighboring community, i.e. the community assigned to one of his connections.

Given the modularity defined by Equation 4.1, we can evaluate the fluctuation resulting from a removal of node i from its community C_i and becoming an isolated node by the following equation:

$$\Delta Q_R(i, C_i) = \left[\frac{l_{C_i} - l_{i,C_i}}{l_G} - \left(\frac{d_{C_i} - d_i}{2l_G}\right)^2 - \left(\frac{d_i}{2l_G}\right)^2\right] - \left[\frac{l_{C_i}}{l_G} - \left(\frac{d_{C_i}}{2l_G}\right)^2\right] \tag{4.2}$$

where l_{i,C_i} is the sum of the weights of the links between the community C_i and node i and d_i is the weighted degree of node i. This computation is very intuitive. Since the changes in the partition are only impacting community C_i, the fluctuation is simply obtained by the difference between the contribution of the new and old community C_i to the modularity function. The old community C_i is the one still containing the node i while the new community is the one without this node.

Similarly, the fluctuation resulting from an insertion of an isolated node i to a neighbouring community C_j is given by the following equation:

$$\Delta Q_I(i, C_j) = \left[\frac{l_{C_j} + l_{i,C_j}}{l_G} - \left(\frac{d_{C_j} + d_i}{2l_G}\right)^2\right] - \left[\frac{l_{C_j}}{l_G} - \left(\frac{d_{C_j}}{2l_G}\right)^2 - \left(\frac{d_i}{2l_G}\right)^2\right] \tag{4.3}$$

The justification of this equation is similar to the one of Equation 4.2. The old community C_j is the one without the node i while the new one contains node i.

The computation of the sensitivity measure can then be described by four main steps explained hereafter.

Computation of the Sensitivity Measure

For each node i in the network:

1. Compute $\Delta Q_R(i, C_i)$, i.e. the fluctuation of the modularity function when removing the node from its assigned community C_i.
2. For each neighboring community C_j of node i, compute $\Delta Q_I(i, C_j)$, i.e. the fluctuation of the modularity function when inserting the node into the neighboring community C_j.
3. Find the neighboring community C_j of node i maximizing the fluctuation $\Delta Q_I(i, C_j)$ from Step 2.
4. Compute the sensitivity measure by adding the fluctuations $\Delta Q_R(i, C_i)$ from Step 1 and $\Delta Q_I(i, C_j)$ from Step 2 where C_j is given by Step 3.

As explained in Step 3 and 4, the sensitivity measure for a node i is then given by

$$S(i) = \Delta Q_R(i, C_i) + \max_{C_j} (\Delta Q_I(i, C_j)) \tag{4.4}$$

$$= \max_{C_j} \frac{1}{l_G} \left[\frac{d_i(d_{C_i} - d_{C_j} - d_i)}{2l_G} - (l_{i,C_i} - l_{i,C_j}) \right] \tag{4.5}$$

One of the main advantages of this method is that it takes advantage of the information computed by the Louvain method. Indeed, the computation of the sensitivity measure $S(i)$ for a single node i is only $O(m)$ where m is the number of different communities. In practice, the value is small and all the variables used in equation 4.5 are already computed by the Louvain method. As a consequence, the incremental complexity of this method is $O(nm)$ where n is the number of nodes since we have to compute the measure for each node of the network.

Applications

Intuitively, the sensitivity measure of a node is the fluctuation of the overall modularity that is produced by its removal from its community and its assignment to the another community. It is thus legitimate to test if reassigning the community of nodes that have a positive sensitivity measure could improve the overall modularity of the partition.

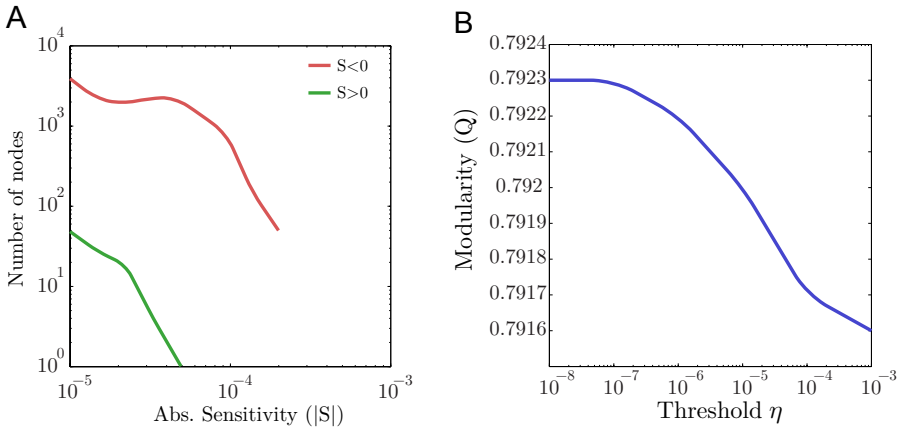


Figure 4.5: **Sensitivity value distributions and impact on modularity.** (A) Distribution of positive (green) and negative (red) sensitivity values for nodes for the tower-based network partition. (B) Evolution of the modularity of the overall partition as nodes with a sensitivity $S > \eta$ are reassigned to their best neighbouring community.

Despite the definition of the sensitivity measure, changing the community of these nodes does not ensure an increase of the modularity of the partition. Indeed, the sensitivity measure evaluates the local fluctuations from a single node. It does not take into account the composition of multiple changes.

To illustrate the feasibility and usefulness of this measure, the sensitivity (S) of each node is computed for the partition of the tower-based network (Fig. 4.2.A). The distribution of sensitivity values reveals that most of the nodes have a negative sensitivity (Fig. 4.5A). This indicates that the partition is rather stable and corroborates its excellent modularity value ($Q = 0.79$) . Nevertheless, some nodes have a positive value ($\approx 1\%$ of nodes) and assigning them to their neighbouring community maximizing S improves the quality of the partition (Fig. 4.5B). Indeed, as we change the community assignment of nodes that have a sensitivity $S > \eta$ we observe an increase of the overall modularity of the partition, as η is lowered and as more nodes are reassigned.

Beyond improving the overall quality of a partition, the sensitivity measure also provides a way to spatially map social interactions strength within a community. The spatial distribution of sensitivity measures of nodes can offer a detailed map of the stability of a particular community, revealing stable and unstable areas as well as quantifying the strength of

the spatial fracture of social interactions between two contiguous regions. To illustrate this, sensitivity values of mobile phone towers are spatially extrapolated for two particular communities corresponding to the administrative region of Alsace (Fig. 4.6.AC) and Champagne-Ardenne (Fig. 4.6.BD). The community identified to Alsace is characterized by a significant social withdrawn as low sensitivity values are observed within the community, i.e. towers within the community are much more connected to each others than to towers located outside. Besides corroborating social studies in that particular region [258, 259], this observation highlights the potential of this measure to map social behaviours of a population in space with high resolution and details.

4.6 Conclusion

For the first time by studying the structure of telecommunications we have the ability to understand the configuration of social communities in France. As observed in numerous studies in the past, interpersonal relationship are driven by geographical proximity and residential propinquity [246–249, 260]. This can also be observed in our constructed networks as about 80% of all calls cover distances of no more than 50km. However, the resulting spatial partitions reveal a surprising regional coherence with interpersonal ties, suggesting that interpersonal relationship seems to be driven as much by administrative boundaries as geographical proximity. It takes no great stretch of the imagination to understand why, as social networks tend to be constructed around geography and driven by institutional propinquity: people meet as school, work, church, social organisations or neighbourhoods. School districts are driven by administrative borders and form the foundation for interpersonal relations, not only among students but among their parents as well due to the fact that administrative boundaries determine where children will receive their education. Additionally, employment patterns play a significant role in the delicate daily balancing act between professional and family obligations and tend to be determined more by administrative boundaries than proximity. The conclusion is that despite technological advances daily interactions continue to revolve primarily around local concerns. Administrative regions which have until now been viewed as no more than a collection of culturally distinct departments would seem to evolve towards a more deeply rooted sense of shared identity.

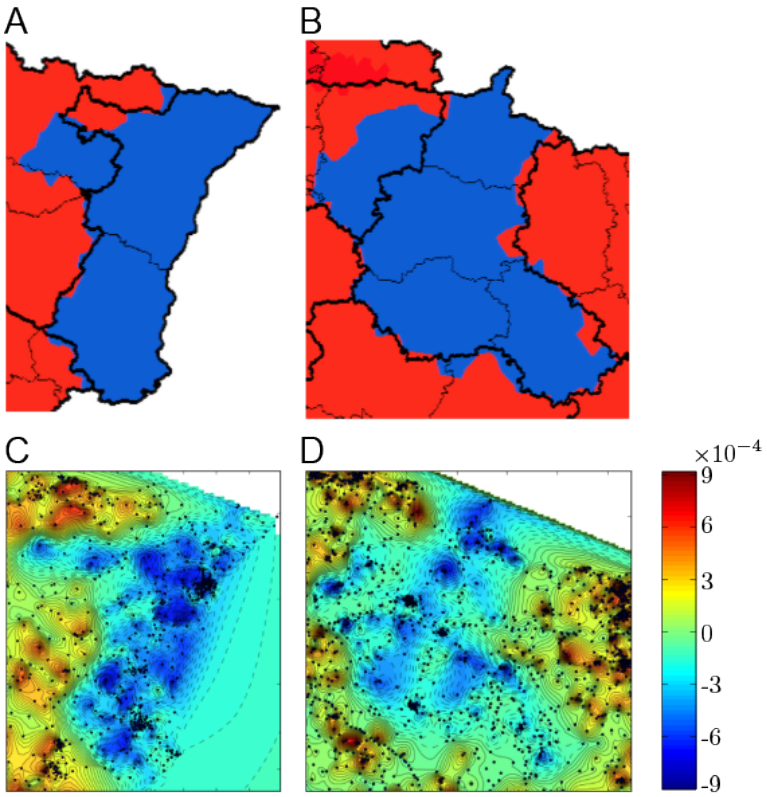


Figure 4.6: **Spatial distribution of sensitivity values. (A)-(B)** Targeted communities (blue) corresponding to the region of (A) Alsace and (B) Champagne-Ardenne. **(C)-(D)** Spatial distribution of sensitivity measure S for nodes (towers) within the targeted communities and $-S$ for outside nodes ((C) Alsace and (D) Champagne-Ardenne). Blue locations within the targeted communities correspond to stable areas as sensitivity values are negative for those nodes. Conversely, red locations outside of the targeted communities corresponds to stable area as well as we take the opposite value of the sensitivity for those nodes.

Beyond these sociological results, network-based tools such as the *sensitivity* measure can reveal with a startling degree of accuracy the stability of small areas and nodes within a particular community. Given its low complexity and intuitiveness, this measure offers promise for extension to other telecommunication networks present in low-income countries and where the understanding of the spatial structure of social communities can be crucial to help tackling the many challenges they are facing.

Besides telecommunication networks, this sensitivity measure can also be computed in many other systems such as citation networks and collaboration networks, suggesting potential opportunities for numerous big data applications, from churn prediction in customer networks to automated document classification in citation networks.

Chapter 5

Connection between social interactions and mobility

In this chapter, we investigate the role space plays on social interactions and human movements and how these two quantities relate to each other. We first describe the three mobile phone datasets used in this study to extract mobility and social fluxes between pair of locations. Second, we derive a scaling relationship between these two quantities that allows us to derive one from the other. Finally, we demonstrate the practical relevance of our results in the context of epidemic spreading.

5.1 Introduction

Over the past few years, we have witnessed tremendous progresses in uncovering patterns behind human mobility [96–98, 114, 140, 261, 262] and social networks [14, 263, 264], owing partly to the increasing availability of and accessibility to large-scale datasets capturing human behavior in a new level of detail, resolution, and scale [65, 265]. These data offer a huge opportunity for research, fueling concomitant advances in both areas of human mobility and social networks with profound consequences in broad domains. One important aspect affecting both areas is the critical role space plays. Indeed, growing evidence suggests both our movements and communication patterns are associated with spatial costs that follow reproducible scaling laws, each characterized by its specific critical components. Indeed, previous studies have shown that human travels adhere to spatial constraints [249], characterized by levy flights and continuous time random walk models [96–98, 112], a scaling law that has proven to be critical in various phenomena driven by human mobility, from disease and pandemic spreading [102, 119, 266, 267] to

migrations [97, 114] and emergency response [100, 268, 269]. In another related yet distinct area, there has been much empirical evidence about the geographic impact on communication patterns [249], documenting that the probability for two individuals to communicate decays with distance, following power law distributions [136, 247–249, 270]. This robust pattern plays an important role in navigating the social network [271], from routing [272, 273] to search of experts [274, 275] to spread of information [136, 276] and innovations [277]. While human movements and social interactions bear high-level similarities in the role spatial distance plays, they remain as largely separate lines of inquiry, lacking any known connections between the two, which is particularly perplexing given the fact that they often exploit the same datasets [112, 140, 249, 278, 279] and are treated indifferently in most modeling frameworks [101, 114].

In this chapter, we test the hypothesis that previously observed spatial dependency captures a convolution of geographical propensity and a popularity based heterogeneity among locations by exploiting three large-scale mobile phone datasets from different countries across two continents. By separating these two factors, we discover a scaling relationship linking the critical exponents associated with the spatial impact on movement and communication patterns, effectively reducing the number of independent parameters characterizing human behavior. The uncovered scaling theory not only allows us to derive human movements from communication volumes, or vice versa, it also hints for a deeper connection that may exist among all networked systems where space plays a role, from transportations [97, 114, 280] and communications [136, 248, 270] to the internet [272, 273] and human brains [281].

5.2 Data description

Mobile communication records, catalogued by mobile phone carriers for billing purposes, provide an extensive proxy of human movements and social interactions at a societal scale. Indeed, by keeping track of each phone call between two users and the spatiotemporal information about the users who initiated and received the call, mobile phone data offers information on both human mobility and social communication patterns at the same time as we will detail hereunder.

In this chapter, we compiled a uniquely rich database consisting of three different datasets that are of a similar level of details yet with

different demographics, economic status, and scales:

D_1 This dataset contains mobile phone calls between 1.3 million users over a period of one month in 2006 from an European country. For each phone call, the caller and the callee, both anonymised with a key (hash code), the time, the date and the phone towers routing the communication are recorded. Only phone calls between users that called each other at least 5 times over a period of 18 month are known. Furthermore, only the coordinates of the mobile phone towers are known, hence the position of a user within the range of an antenna is unknown.

D_2 This dataset covers a six-month period of mobile phone calls between 6 million anonymised users from a large European country. For each phone call, the caller, the callee, the time and the towers routing the communication are recorded. Similarly to D_1, only the coordinates of the mobile towers are known, hence the position of a user within the range of an antenna is unknown.

D_3 The dataset covers a period from 2005 to 2009 and is made of all transaction logs of all mobile phone activity that occurred in an African country over the 5 year-period. The data originate from the largest mobile phone operator in that country and contain about 1.5 million phone calls. The logs include the date, the time, and the mobile phone towers routing the call for each of the phone calls and are again anonymous. Again, only the coordinates of the mobile towers are known, hence the position of a user within the range of an antenna is unknown.

For each of these datasets, mobility and socials fluxes capturing the movements of customers and their social communications between pairs of locations can be inferred from phone call information:

Mobility fluxes For each phone call, the position of the tower routing the call is known for the caller. Since we know the location of each tower, we know the location of the user was within the range of the tower's service area. By looking at each consecutive phone calls made by a user, we can thus reconstruct the user's jumps between two consecutive locations where his calls were initiated. By aggregating all movements for all users, we can thus obtain the total number of jumps from any

tower i to any tower j ($T^M_{i,j}$). All jumps made outside continental territories (i.e. islands) were not taken into account. The jumps do not exceed $\sim$ 1000 km, $\sim$ 400 km and $\sim$ 100 km for datasets D_1, D_2, D_3, respectively, due to national frontiers and coverage limitations driven by geographical constraints in the country. We consider the number of jumps between two locations as the mobility fluxes between them.

Social fluxes For each phone call, the position of the tower routing the call is known for both the caller and the callee. By considering all phone calls, we thus know the total number of calls from a tower i to a tower j ($T^S_{i,j}$). We consider the number of phone calls between two locations as the social fluxes between them.

5.3 Scaling relationship

To quantify the spatial effect on social communication patterns, we often measure the distance distribution of communications using two oft-used distance metrics:

Communication distance distribution: The distance r characterizing social communications is the geodesic distance between two individuals u and v, who communicate via phone calls or SMS. Previous studies suggested that the probability for two individuals to communicate decreases with distance, following a power law distribution [24, 248, 249]. Here we recovered previous results (Fig. 5.1A), finding that the distance distribution of each studied system, $P^S(r)$, can be approximated as:

$$P^S(r) \sim r^{-\beta^r_i}. \tag{5.1}$$

We find, among different countries, the exponents β^r_i are characterized by rather small variations ($\beta^r_i \approx 1.5$) (Fig. 5.1A, Table 1).

Rank distribution: Within a country, the populations are not distributed uniformly in space. To account for such inhomogeneity, previous studies proposed the rank measure as an alternative means to quantifying the effective distance between two individuals [136]. The rank between two users u and v is the number of people closer to u than v, formally defined as $s = |w : r(u, w) < r(u, v)|$. We measure the rank distributions for our three datasets (Fig. 5.1B), finding $P^S(s)$ is characterized by a power law tail, consistent with previous studies [136,

249]:

$$P^S(s) \sim s^{-\beta_i^s}. \tag{5.2}$$

The exponents β_i^s for our three datasets are shown in Table 1.

Similarly, for human movements, the jump size distribution is most commonly used to quantify spatial constraints in human movements. Here we measure this quantity in different distance metrics :

Jump size distribution: Jump size measures the displacement in the unit of kilometers between two consecutive sightings of an individual. A fundamental property of human mobility is that the aggregated jump-size distribution follows a power law [96–98],

$$P^M(r) \sim r^{-\alpha_i^r}, \tag{5.3}$$

indicating most of the time people travel over short distances, between home and work for example, while they occasionally take longer trips. We measured $P^M(r)$ in our data corpus (Fig. 5.1C), finding few variations in α_i^r among different countries ($\alpha_i^r \approx 1.9$).

Rank jump-size distribution: To account for biases from population density we measure the rank s of each jump. We find that $P^M(s)$ is also characterized by a power law tail as suggested by previous studies [112, 249],

$$P^M(s) \sim s^{-\alpha_i^s}. \tag{5.4}$$

As shown in Fig. 5.1D, α_i^s is similar for D1 and D2 ($\alpha_i^s \approx 1.3$) but different from D3: $\alpha_3^s \approx 1$ (Table 1).

Taken together, the spatial scaling of social interactions ($P^S(r)$ and $P^S(s)$) for dataset i is characterized by exponents β_i^r and β_i^s, respectively, while human movements ($P^M(r)$ and $P^M(s)$) by exponents α_i^r and α_i^s. These quantities were reported previously by independent research groups with different measurement details [96, 97, 248, 282]. Here we measure these quantities systematically by using a comprehensive database we compiled. We find that, within each of the two categories, the critical exponents (α_i or β_i) in different countries are rather similar to each other. For example, there is little difference between the three α_i^r or β_i^r exponents. For the rank metrics, D1 and D2 are also very similar to each other, while D3 is characterized by slightly different exponents. Yet, most noticeably, we observed substantial and systematic differences between $\alpha_i^{r,s}$ and $\beta_i^{r,s}$. Such differences contradict current modelling frameworks from gravity model [282] to radiation model [114]

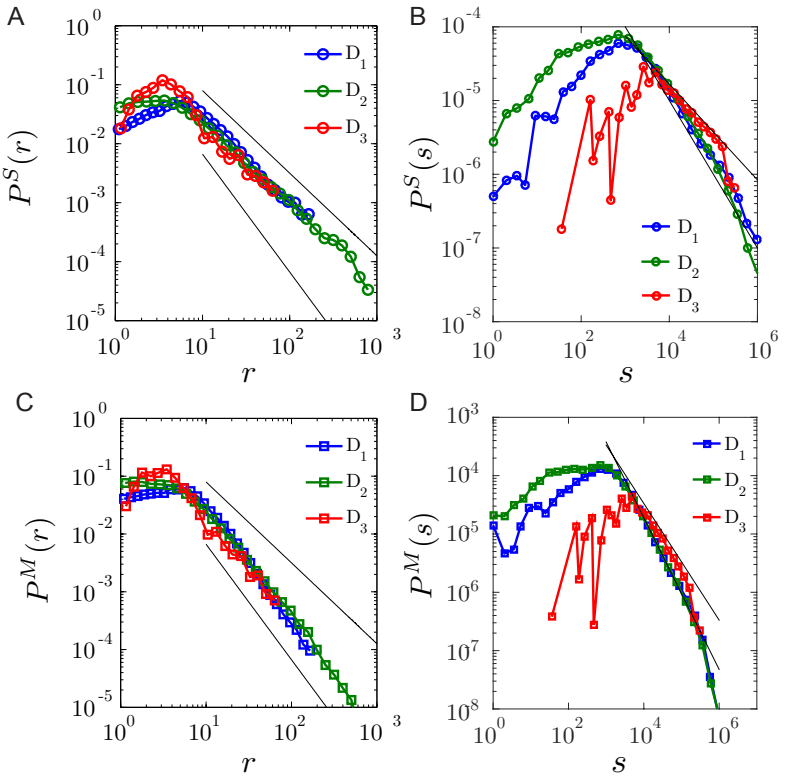


Figure 5.1: **Mobility and communication distance distributions.** **(A)** Communication distance distributions measured in geodesic distance r, $P^S(r)$, for all three datasets follow a power law distribution with exponents $\beta^r \approx 1.5$. **(B)** Rank distributions $P^S(s)$ for the three datasets follow a power law distribution with exponents $\beta^s = 1$ for D_1 and D_2 and $\beta^s = 0.65$ for D_3. **(C)** Jump-size distribution $P^M(r)$ measured in geodesic distance r follows a power law distribution with exponent $\alpha^r \approx 1.9$. **(D)** Rank jump-size distribution $P^M(s)$ for rank s follows a power law distribution with exponent $\alpha^s \approx 1.3$ for D_1 and D_2 and $\alpha^s \approx 1$ for D_3.

that treat these two classes of problems as the same phenomena given a population distribution, thus predicting the same scaling exponent within each country [24, 248]. This raises a critical question: What is the origin of the observed differences between exponents α_i and β_i?

$P^S(s)$ (or $P^S(r)$) measures the intensity of social communications as a function of distance, capturing on a population averaged level the social fluxes between different locations. On the other hand, $P^M(s)$ (or $P^M(r)$) measures the aggregated jumps between places, corresponding to the mobility fluxes from one location to another. Denoting with $T^S_{i,j}$ the social fluxes from location i to j and with $T^M_{i,j}$ the mobility fluxes, i.e., the total number of communications/jumps between two locations, we measure $T^S_{i,j}$ and $T^M_{i,j}$ between any two locations over a one month period. We find that both social and mobility fluxes follow fat-tailed distributions across our three studied datasets (Fig. 5.2). This is somewhat expected: Indeed, if we view each location as a node and fluxes as links connecting different locations, the fat-tailed distributions of fluxes are consistent with previous results on link weight distributions [283]. Hence, Fig. 5.2 documents an inherent heterogeneity between locations. Indeed, there are few fluxes between most locations, yet a non-negligible fraction of location pairs are characterized by a large number of fluxes. The fat-tailed nature of flux distributions raises an important question: Can distance dependencies (Fig. 5.1) be accounted for by the observed heterogeneity in fluxes alone (Fig. 5.3)? To this end, we take D1 as an exemplary case and control for spatial effect by choosing location pairs that are of similar distances (s), and measuring the distributions for social ($P^S_T(T \mid s)$ in Fig. 5.3A) and mobility fluxes ($P^M_T(T \mid s)$ in Fig. 5.3B), respectively. We find the fluxes still follow a fat-tailed distribution within each group, indicating there still exists much heterogeneity in fluxes even among locations within similar distances. Moreover, locations that are nearby (small s) tend to have higher fluxes, corresponding to higher intensity in both communications (Fig. 5.3A) and movements (Fig. 5.3B). Indeed, the curves in Fig. 5.3AB shift to the right as s decreases, indicating the probability for two locations to have large fluxes decays with distance. This is consistent with preceding results (Fig. 5.1, Eq. 5.1 and Eq. 5.3), as most communications and movements take place in short distances, accounting for majority of the fluxes. Yet, as shown in Fig. 5.3AB, not all pairs of nearby locations have large fluxes. To the contrary, most of them have very few fluxes. Rather, it is a small fraction of location pairs in each distance groups, i.e., the tails of $P^S_T(T \mid s)$ and $P^M_T(T \mid s)$, that are responsible for generating the majority of fluxes. Most surprisingly, once we rescale the flux distributions with the aver-

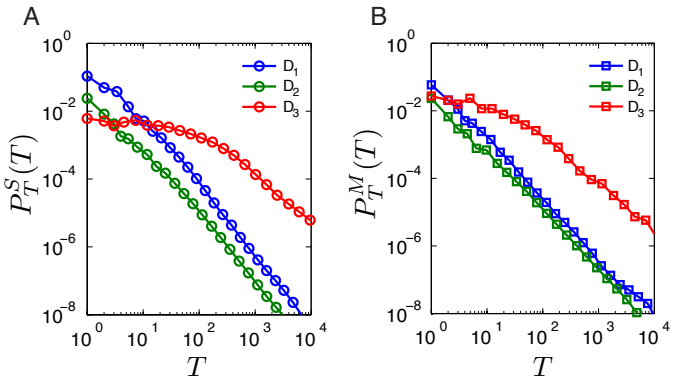


Figure 5.2: **Social and mobility flux distributions.** Distribution of **(A)** social fluxes T^S and **(B)** mobility fluxes T^M for all three datasets.

age fluxes, $\left\langle T^S(s)\right\rangle$ or $\left\langle T^M(s)\right\rangle$, we find all the curves shown in both Fig. 5.3A and Fig. 5.3B (10 curves in total) collapse into one single curve, suggesting a single universal flux distribution characterizes both social interactions and human movements, independent of distance (Fig. 5.3C).

This data collapse indicates that

$$P_T^{S,M}(T \mid s) = \left\langle T^{S,M}(s)\right\rangle^{-1} \mathcal{F}\left(T^{S,M} / \left\langle T^{S,M}(s)\right\rangle\right), \qquad (5.5)$$

where $\mathcal{F}(x)$ is a distance-independent function. The data collapse in Fig. 5.3C is rather remarkable. It indicates that the observed localization in social communications and human movements can be decomposed into two independent factors: one is the universal distribution $\mathcal{F}(x)$ which is distance independent, characterizing the inherent popularity-based heterogeneity among different locations. All the distance dependencies are now encoded in the average fluxes at a given distance, i.e., $\left\langle T^S(s)\right\rangle$ for social and $\left\langle T^M(s)\right\rangle$for mobility fluxes. We repeated our measurements using r as distance metric as well, finding again an excellent data collapse (Fig. 5.3D-F). We also repeated our analysis for datasets D_2 and D_3 and found consistent results as all curves collapse into one for both geodesic (r) and rank (s) distances, demonstrating the robustness of our findings (Fig. 5.4).

The uncovered universal function indicates that the social and mobility fluxes are important factors to characterize communication and mobility patterns, prompting us to measure correlations between the two quantities. We group location pairs (i and j) based on their distance and measure the relationship between $T^S_{i\to j}(s)$ and $T^M_{i\to j}(s)$ for each

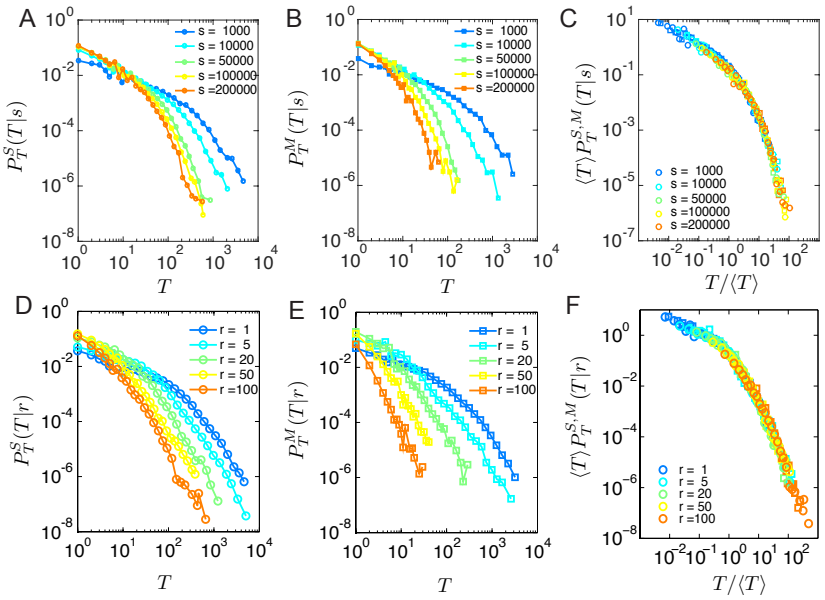


Figure 5.3: **Collapse of social and mobility flux distributions for D_1. (A)** Flux distributions of communication for different rank groups, $P^S(T|s)$. **(B)** Same distributions as (A) for mobility fluxes, $P^M(T|s)$. **(C)** Mobility and communication fluxes, denoted by circles and squares, respectively, collapse into one single curve after rescaled by the average fluxes in each group $\langle T\rangle$. **(D)** Flux distributions of communication for different distance groups, $P^S(T|r)$ (same as (A) but measured in geodesic distance r). **(E)** Same distributions as (D) for mobility fluxes, $P^M(T|r)$. **(F)** Mobility and communication fluxes measured in geodesic distance, denoted by circles and squares, respectively, again collapse into one single curve after rescaled by $\langle T\rangle$ for the different geodesic distance groups.

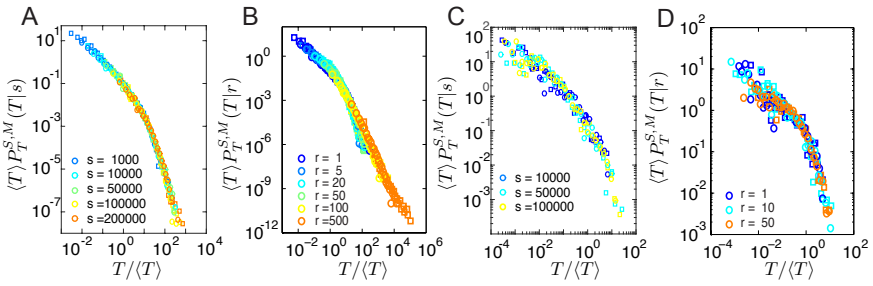


Figure 5.4: **Collapse of social and mobility flux distributions for D_2 and D_3. (A)** Mobility and communication fluxes distributions for D_2, denoted by circles and squares, respectively, collapse into one single curve after rescaled by the average fluxes in each group $\langle T\rangle$. **(B)** Same distributions as (A) but for geodesic distances. **(C)-(D)** Same distributions as (A)-(B) but for dataset D_3.

group ($s = 1e3$, $s = 1e4$, $s = 5e5$, $s = 1e6$, and $s = 2e6$ in Fig. 5.5A-E). In these scatter plots, each grey dot represents a pair of locations, and its x-y coordinates correspond to the mobility ($T^M_{i\to j}(s)$) and social ($T^S_{i\to j}(s)$) fluxes from i to j. We find strong correlations between these two quantities regardless of how faraway these locations are separated. To quantify this correlation, we measure the average social fluxes given the mobility fluxes at a certain distance, $\overline{T^S}(T^M|s)$ (colored symbols in Fig. 5.5A-E), which is formally defined as

$$\overline{T^S}(T^M|s) \equiv \frac{\sum_{i\to j} T^S_{i\to j}\delta(T - T^M_{i\to j})\delta(s - s_{ij})}{\sum_{i\to j}\delta(T - T^M_{i\to j})\delta(s - s_{ij})}, \tag{5.6}$$

where $\delta(x)$ is the delta function ($\delta(x) = 1$ when $x = 0$, and $\delta(x) = 0$ otherwise). We find in Fig. 5.5A-E that the average social fluxes $\overline{T^S}(T^M|s)$ has a scaling relationship characterised by a slope of $\theta_s = 0.9$, indicating social fluxes scale sub-linearly with mobility fluxes, independent of distance s. The shift along the y-axis through Fig. 5.5A to E also reveals the existence of a pre-factor $A(s)$. We find indeed, as distance increases, that the average social fluxes increases given the same volume of mobility fluxes. Hence, $A(s)$ characterises the cost tradeoff between phone communications and commuting. As a result, $\overline{T^S}(T^M|s)$ has a power law scaling relationship with T^M of the form

$$\overline{T^S}(T^M|s) = A(s)T^M(s)^{\theta_s}, \tag{5.7}$$

where the scaling exponent $\theta_s = 0.9$ for different s. Rescaling $\overline{T^S}$ by s^{δ_s}, we find all curves collapse into a straight line (Fig. 5.5F), indicating

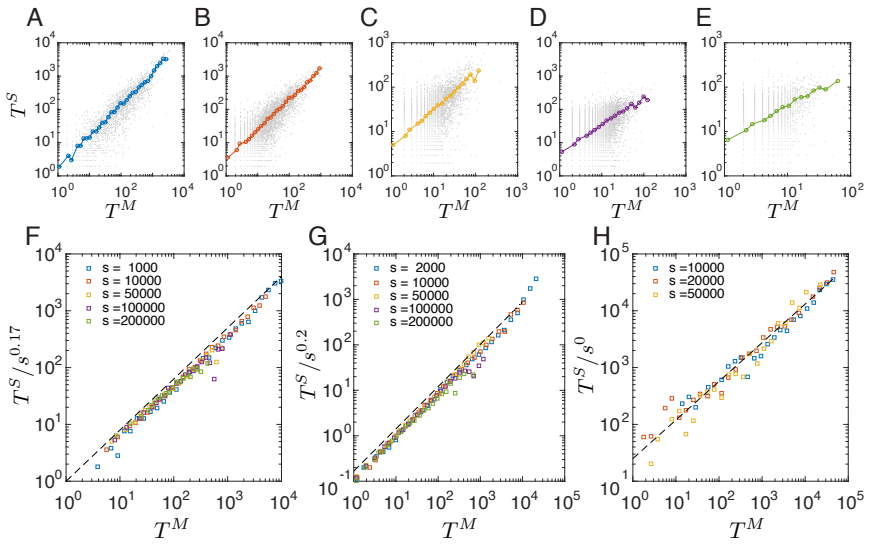


Figure 5.5: **Correlation between social and mobility fluxes using rank distance.** Correlations between $T^S_{i\to j}(s)$ and $T^M_{i\to j}(s)$ for location pairs (grey dots) separated by a distance of **(A)** $s = 1e3$, **(B)** $s = 1e4$, **(C)**$s = 5e5$, **(D)** $s = 1e6$, and **(E)** $s = 2e6$. We find all curves collapse into a straight line when $\overline{T^S}$ is rescaled by s^{δ_s} for **(F)** D_1, **(G)** D_2, and **(H)** D_3.

$A(s) \sim s^{\delta_s}$ where $\delta_s = 0.15$. We repeated the same measurement for $D2$ and $D3$. We found, although each dataset is characterized by a different set of θ_s and δ_s, Eq. 5.7 holds consistently well across different datasets (Fig. 5.5GH). We also repeated our analysis by replacing s with other distance metrics (geodesic distance r), finding again consistent results with Eq. 5.7 (Fig. 5.6). Indeed, each dataset is well described by its characteristic set of θ_r and δ_r exponents, demonstrating the robustness of our findings.

Most important, Eq. 5.7 together with the data collapses in Fig. 5.3CF (Eq. 5.5) allows us to derive a new scaling relationship between different critical exponents. Indeed, the average social fluxes at distance s, $\overline{T^S}(s)$, can be obtained by integrating $\overline{T^S}(T^M, s)$ over T^M:

$$\overline{T^S}(s) = \int P_T^M(T^M|s)\overline{T^S}(T^M, s)dT^M. \tag{5.8}$$

Substituting Eqs. (5.5) and (5.7) into (5.8), we have

$$\overline{T^S}(s) = \int \mathcal{F}(x)\overline{T^S_M}\left(\overline{T^M}(s)x, s\right)dx \sim \overline{T^M}(s)^{\theta_s}s^{\delta_s}\int x^{\theta_s}\mathcal{F}(x)dx, \tag{5.9}$$

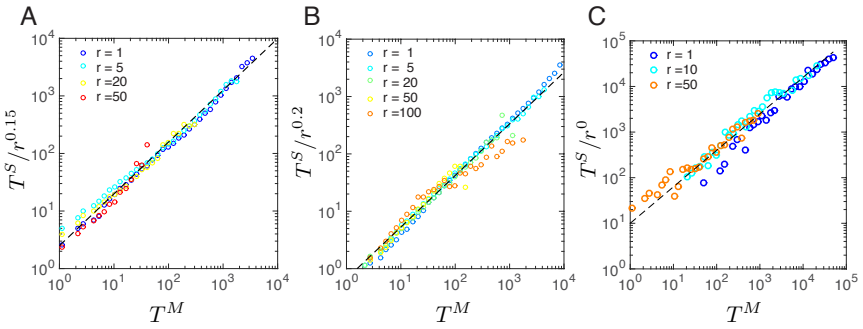


Figure 5.6: **Correlation between social and mobility fluxes using geodesic distance.** Correlations between $\overline{T^S}$ rescaled by r^{δ_r} and T^M. Similarly to Fig. 5.5, we find all curves collapse into a straight line when $\overline{T^S}$ is rescaled by r^{δ_r} for **(A)** D_1, **(B)** D_2, and **(C)** D_3.

where $x \equiv T^M/\overline{T^M}$ as a change of variable. As $\overline{T^S}(s) \sim \sum_{i\to j} T^S_{i\to j}\delta(s - s_{ij}) = P^S(s) \sim s^{-\beta_s}$, and similarly $\overline{T^M}(s) \sim s^{-\alpha_s}$, we have,

$$s^{-\beta_s} = s^{-\alpha_s\theta_s} s^{\delta_s} \int x^{\theta_s} \mathcal{F}(x) dx. \tag{5.10}$$

The tail behavior of $\mathcal{F}(x)$ indicates the integral in Eq. 5.10 converges. Hence, Eq. 5.10 leads to a scaling relationship:

$$\beta_s = \alpha_s\theta_s - \delta_s, \tag{5.11}$$

connecting the exponent that characterizes social communications (β_s) and the exponent characterizing human movements (α_s). Similarly, for geodesic distance metric r, we obtain:

$$\beta_r = \alpha_r\theta_r - \delta_r. \tag{5.12}$$

We measure each exponent in Eq. 5.11 and Eq. 5.12 independently for different datasets, finding excellent agreement between empirical measurements and our theoretical predictions (Table 1). Hence, Eq. 5.11 and 5.12 offer an explicit link between critical exponents characterizing spatial dependencies in human movements and social interactions, showing that the social exponent (β) can be expressed in terms of the mobility exponents (α), a consistently robust result that is independent of distance metrics being used. The uncovered scaling relationship between these two classes of exponents is mediated by a universal flux distribution ($\mathcal{F}(x)$) we uncovered in this study. This scaling relationship

Table 5.1: Critical exponents. We measured α_s, β_s, θ_s, and δ_s independently for each dataset by using rank as distance metric. We then compute $\tilde{\beta}_s = \alpha_s\theta_s - \delta_s$ using Eq. 5.11, finding $\tilde{\beta}_s$ largely agrees with β_s across different datasets. Similarly we repeated the same measurements by using geodesic distance, obtaining α_r, β_r, θ_r, and δ_r and hence computing $\tilde{\beta}_r$. We find $\tilde{\beta}_r$ also well approximates β_r, with only exception observed in D_3 ($\Delta \sim 0.13$), which is likely due to its smaller data size. As both our data size and non-integer nature of distance metrics prevents us from using standard fitting algorithms for power laws [284], we computed all our exponents by using the least-square method [285, 286].

	α_r	α_s	β_r	β_s	θ_r	θ_s	δ_r	δ_s
D_1	2	1.3	1.6	1	0.9	0.9	0.2	0.15
D_2	1.9	1.3	1.5	1	0.9	0.93	0.2	0.2
D_3	1.9	1	1.5	0.65	0.8	0.68	0	0

bridges two fields that are largely pursued disjointly [249], showing that they represent different facets of a deeper underlying reality, effectively reducing the number of independent parameters characterizing human behavior. Next we show the uncovered relationship offers us a powerful framework to derive quantities pertaining to one field from those of the other.

5.4 Application to spreading processes

We simulate a virus spreading process using D1 as an exemplary case to demonstrate how the above findings can be used to connect human mobility and social interactions in a practical context. Of the many ingredients in computational modeling of virus spreading, human mobility is among the most critical [22, 96, 102, 266, 287, 288]. To understand how human movements catalyze societal-wide spreading processes, we infect a few randomly selected individuals with some hypothetical germ in a random location at time $t = 0$. Denoting with μ the infection rate of this germ, we assume that, at each time step, an infected individual could spread the disease to others within his/her vicinity, i.e., individuals within the same mobile tower. At the same time, any infected individual can recover from the disease at rate ν. This process is known as the Susceptible-Infected-Susceptible (SIS) model, commonly used in modeling disease spreading [289, 290].

Choosing any set of μ and ν, we can simulate a spatial SIS model by following the real mobility fluxes between locations ($T^M_{i,j}$) measured from our dataset. This raises an interesting question: had we not had access to mobility information, how well can we approximate the observed spreading pattern using the social fluxes rescaled by our scaling relationship uncovered in Eq. 5.11?

Following Eq. 5.7 and using the exponents from Eq. 5.11, mobility fluxes between a location i and j can be approximated by rescaled social fluxes, $\tilde{T}^S_{i,j}$, defined as

$$\tilde{T}^S_{i,j} = (s^{-\delta_s}_{i,j} T^S_{i,j})^{-\theta_s}. \tag{5.13}$$

where $\delta_s = 0.15$, $\theta_s = 0.9$ for D1 (Table 5.1) and $s_{i,j}$ is the distance between the two locations. We simulate a spreading process in Portugal using the real mobility fluxes T^M, the rescaled social fluxes $\tilde{T}^S$ as well as the mobility fluxes T^M_{GM} approximated by the widely used gravity model [103, 105, 249, 282]. To compare these results, we started from the same initial conditions ($\mu = 0.9$, $\nu = 0.3$) and initial infected users located in Lisbon are used for all three simulations in this example.

Denoting with n_i the number of users at location i, with a_i the area of location i, and $m_i(t)$, $\tilde{m}_i(t)$ and $m^{GM}_i(t)$ the number of infected users at time t in location i when using T^M, $\tilde{T}^S$ and T^M_{GM}, respectively, we measure $m_i(t)/a_i$, $\tilde{m}_i(t)/a_i$ and $m^{GM}_i(t)/a_i$, i.e. the density of infected users estimated in each location i for the three cases (Fig. 5.7ABC for $t = 17$). We find a remarkable agreement between our simulation and the real spreading patterns. Moreover, close up on the city of Porto reveals a superior accuracy of our model comparing with predictions from gravity model. To quantify the differences between the two methods, we measure

$$\tilde{e}_i(t) = \frac{m_{i,t} - \tilde{m}_{i,t}}{m_{i,t}} \tag{5.14}$$

and

$$e^{GM}_i(t) = \frac{m_{i,t} - m^{GM}_{i,t}}{m_{i,t}} \tag{5.15}$$

corresponding to the relative error of infection rate in each location i at time t for both methods (Fig. 5.7DE at $t = 17$). The drastic difference between Fig. 5.7D and Fig. 5.7E highlights the fact that lower $\tilde{e}_i(t)$ are observed comparing with $e^{GM}_i(t)$ in this particular example, again documenting the superior predictive power of our model.

In order to systematically assess and compare the accuracy of our results, we simulated 500 independent spreading processes following the

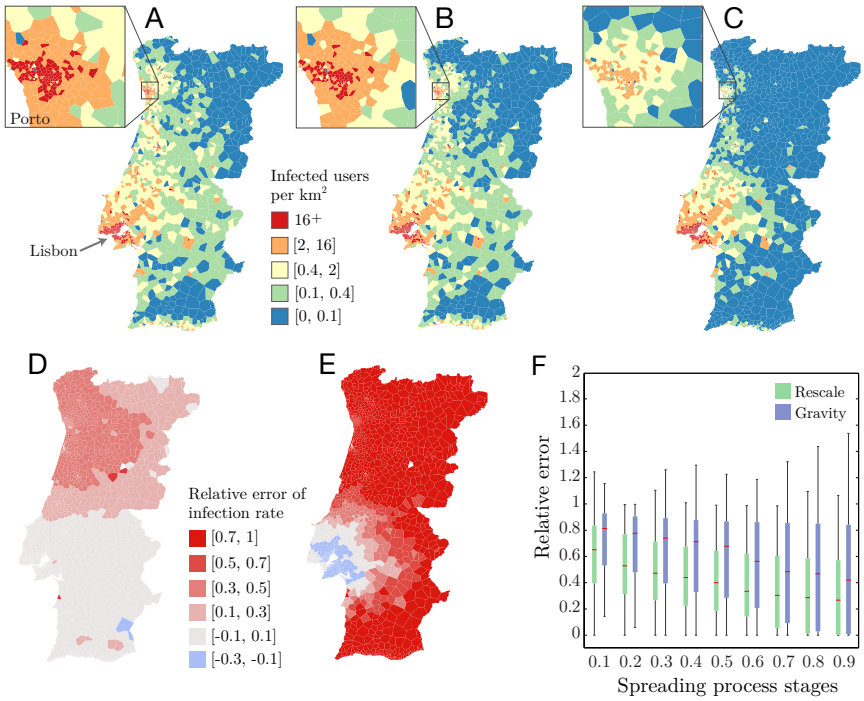


Figure 5.7: **Simulation of SIS spreading processes.** **(A-C)** Densities of infected users at time $t = 17$ following a simulation of a SIS spreading process ($\mu = 0.9$, $\nu = 0.3$) originated from Lisbon by using **(A)** real mobility fluxes T^M, **(B)** the rescaled social fluxes $\tilde{T}^S$, and **(C)** the mobility fluxes approximated by gravity model T^M_{gm}. **(D)** Relative errors of infection rate $\tilde{e}_i(t)$ and **(E)** $\tilde{e}^{gm}_i(t)$ at each location i at time $t = 17$. **(F)** Distributions of mean relative error $\overline{\tilde{e}}(t)$ (green) and $\overline{e^{gm}}(t)$ (blue) over 500 SIS simulations at different stages before reaching the steady state, documenting the superior predictive power of our method comparing with gravity model at all stages of the spreading processes

same procedure described above but choosing randomly parameters μ and ν as well as the initial infected location and the number of infected users. For each simulation, we compute the mean values $\overline{\tilde{e}}(t)$ and $\overline{e^{GM}}(t)$ from Eqs. 5.14 and 5.15 respectively, at different stages (time steps). We find that $\overline{\tilde{e}}(t)$ obtained from the 500 simulations are systematically lower than $\overline{e^{GM}}(t)$ across all stages of the spreading processes (Fig. 5.7F), demonstrating the practical relevance of our scaling relationship that effectively predicts mobility patterns using social communications.

5.5 Conclusion

Taken together, by analysing three large-scale mobile phone datasets from three different countries, we uncovered a new scaling relationship between the critical exponents that characterise spatial dependencies in human mobility and social interactions. This scaling relationship is mediated by a universal flux distribution for both movement and communication patterns, indicating the previously observed distance dependencies capture a convolution of geographical propensity and a popularity based heterogeneity among locations. Separating these two factors allows us to establish a formal connection between different critical exponents that were perceived as independent. We offered theoretical basis for the uncovered scaling relationship, which is further supported by extensive simulations that not only demonstrate its practical relevance but also provide empirical validation of our approach. While we uncovered this relationship using two distance measures independently, we did not study their interdependence. However, investigating the underlying connection between rank-based and geodesic distances could be a research worth of interest as this might help us understand the systematic differences observed between exponents α^r and α^s and between β^r and β^s.

Together our results document a new order of regularity that helps deepen our quantitative understanding of human behavior. Lastly, our results may reach far beyond communications and transportations studied in this paper, as many networked systems are also subject to spatial costs in establishing connections in a very similar fashion as our quoted examples, from routers linked by physical cables to form globally connected internet to axons that connect different regions of human brains. Hence our results may provide relevant insights to a diverse set of networked systems where space plays a role [249], opening up a promising

direction for future investigation.

Part II

Social mechanisms of success

Chapter 6

Information retrieval in large-scale publication data

Besides phone call data, publication data represent another valuable source of information to study social systems. However, the extraction of meaningful data from publication records is far from trivial as many ambiguities can be present. In this chapter, we present three distinct techniques that aim at resolving these issues. In the first part of this chapter, we address the problem of author name disambiguation by presenting an agglomerative method that can automatically merge articles associated to a same author. Following this method, we introduce a technique to disambiguate geographical traces present in publication data. We close this chapter by presenting a novel approach to identify publications related to a same particular topic or field.

6.1 Introduction

Motivation Over the past few years, we have witnessed rapid advances in our understanding of the modern scientific enterprise, owing partly to the massive growth of global research activity as well as the increasing availability of large-scale datasets that capture scientific outputs. These data offer a tremendous opportunity for research on social dynamics, providing results with profound consequences in broads domains. Indeed, recent studies helped us uncover basic mechanisms that govern not only scientific impact [55, 291, 292] but also collaborations [39, 40], credit allocation [43–46] and reputation [293, 294], providing policy-makers effective tools to evaluate in a better and more transparent way scientific outputs. In another related area, research on human mobility exploited the geographical digital traces present in these data,

revealing the global migration of science [295] and uncovering the effect of geography on its dynamic [41, 42].

One common critical aspect about these studies is the importance of acquiring comprehensive disambiguated data. For instance, research on collaboration and reputation requires disambiguated sets of author names [39, 296]. Indeed, as an author's identity is usually represented by a name string in the raw data, name ambiguity can easily appear between individuals sharing the same name, leading to false findings about fundamental characteristics and erroneous predictions [297, 298]. In a similar way, ambiguities also emerge in geographic name entities, which must be taken care of when exploring individual career trajectories in scientific mobility studies [41, 42]. Finally, research on information networks investigating the emergence of ideas or trends also often requires topic-disambiguated data [299, 300].

In this chapter, we address the issues related to data ambiguity by presenting several disambiguation techniques. First, we propose an agglomerative approach to deal with author name redundancy and to detect unique authors. Next, we introduce a geo-tagged and agglomerative method that can efficiently resolve ambiguities in affiliation names. Finally, we present a novel approach that can automatically detect topic-related articles based on their citation network.

Data Two distinct datasets are used to quantify the efficiency of our developed approaches: the *American Physical Society* and *Web of Science*^{TM} datasets.

The dataset provided by the *American Physical Society* (APS) [301] consists of all the papers published in Physical Review, spanning across 9 different journals: Physical Review A, B, C, D, E, I, L, ST and Review of Modern Physics, from 1893 to 2010, amounting to over $450,000$ publications. For each paper the dataset includes title, date of publication (day, month, year), names and affiliations of every author, and a list of the previous APS papers cited.

The dataset provided by the *Web of Science*^{TM} includes several types of scientific outputs such as articles, letters, reviews, editorials and abstracts from 1898 to 2013 across more than 22,000 scientific journals from broad domains, resulting in a set of more than 50 millions papers. For each paper, the dataset includes more than 100 types of information such as the date of publication (month, day, year), the journal issue,

the references towards past items within the dataset as well as author names and affiliations.

6.2 Author Name disambiguation

Identifying authors of an article and, conversely, identifying all articles belonging to a single individual is a fundamental problem. While identifying author names would seem to be a simple process, it represents a major, yet unsolved problem for information science. In this section, we give a short overview of the different problems associated to this issue as well as the main types of solutions that have been proposed over the last decades. Based on these solutions, we then present a simple author name disambiguation technique that can be easily applied on our dataset of interest and which still provides accurate disambiguated results.

Overview
The task of disambiguating author names is associated to four distinct and well-known challenges [302]: (i) individuals may publish under different names, due to spelling variants, spelling errors or pen names for example, (ii) many individuals share the same name, e.g. chinese names [303], (iii) the proportion of interdisciplinary and multi-institutional articles, which are hard to disambiguate, is quickly increasing [304], and finally (iv) the necessary metadata such as citations or affiliations are often incomplete or missing entirely.

If this task was under appreciated in the past, this is far from being the case today. As the internet rose two decades ago as well as massive digital libraries, numerous platforms as well as publishers are now shifting their focus on the search of individuals and not topics or keywords anymore, putting their efforts towards the disambiguation of individuals. If manual disambiguation methods based on collaborative efforts were used in the past, this is not possible today anymore; Thousands of new articles are published everyday in about 28,000 thousands journals, most of them being collected by major publishing platforms. Massive name disambiguation tasks are thus needed and such methods, which heavily relies on individuals efforts, are not feasible anymore.

Under these circumstances numerous approaches have been proposed over the past few years. Overall, these methods can be divided into two different groups: *author grouping* methods [305–308] and *author*

assignment methods [309–312].

The main idea of *author grouping* methods is to rely on a similarity function on the attributes of two articles to decide whether to group them or not. The goal is thus to obtain a function that returns a high similarity for articles authored by the same individuals but a low similarity for articles authored by different ones. As a result, articles corresponding to a same individual will be grouped together, maximising intra- and minimising inter-group similarities. In the literature, there exist two distinct ways to define this similarity function. The first, which is unsupervised, is to rely on predefined functions that compute similarities between given attributes [307, 308]. Such functions includes the cosine similarity, levenshtein distance or tf-idf measure [313]. The second option, which is supervised, is to learn the similarity function through a machine learning algorithm [305, 306]. While learning usually produces better results, it also requires a training sample of disambiguated articles specific for the task, which is often not available. The first unsupervised option, however, does not need such training samples and can thus be more easily adapted to the specific dataset and disambiguation task.

Author Assignment Methods, on the other hand, do not merge articles together but directly assign an article to a given author by constructing a model that represents the author. Such model often includes the probabilities of the author to publish with particular co-authors, in particular venues or on specific topics. For this type of methods, two different techniques exist: *supervised classification* technique [309, 310] and *model-based clustering* technique [311, 312]. The classification technique assigns articles to a particular author using a supervised machine learning algorithm. Based on articles features as well as author features extracted from a training dataset, the algorithm builds relationship between articles and authors features. Ambiguous articles are then divided into distinct sets, each associated to a single individual, in accordance with these relationships. Similarly to the supervised approach of the *author grouping* methods, this technique also requires the acquisition of disambiguated data that requires skilled human annotators. On the other hand, *model-based clustering* technique do not require any training data. Indeed, this technique uses an iterative probabilistic approach to automatically determine the relationship between author and articles features. Even though no training data is needed, this technique still requires specific information about the number of authors in the dataset or the number of author groups, i.e. groups of authors that publish together, which are often not known or hard to estimate.

As we can notice, most of these methods are either based on training data, which requires massive manually disambiguated data, or either very specific to the type of input data, i.e. the metadata available or the nature of the data. Given the large-scale nature of our research, such prerequisites prevent us from implementing or using directly most of these methods. As a result, we present here a simple, yet adaptive method that can efficiently disambiguates large-scale publication data and which relies only on common metadata. This approach could be classified as an unsupervised *author grouping* methods as it does not require any training data and group articles together based on their similarity.

Agglomerative approach
For this method, we consider each author of each publication to be a unique one (number of papers times average number of authors per paper). The intuition behind the disambiguation process is to reduce this number by "merging" authors iteratively, based on a list of criteria. That is, for two publications that were thought to belong to two distinct authors, we iteratively consider them to belong to the same individual if they fullfill all the following:

1. Last names of the two authors are identical;
2. Initials of the first names and, when available, given names are the same. If the full first names and given names are present for both authors, they have to be identical;
3. One of the following is true:
    - The two authors cited each other for at least once;
    - The two authors share at least one co-author;
    - The two authors share at least one similar affiliations (measured by cosine similarity and tf-idf metrics) [314].

The process stops when there is no pair of authors to merge.

Results on the American Physical Society dataset
Of all the publications within the APS dataset, we consider only those for which: (i) there is no ambiguity between an author and his/her affiliation (an ambiguity is present when more than one author and more than one affiliation are given without any link between them); (ii) there

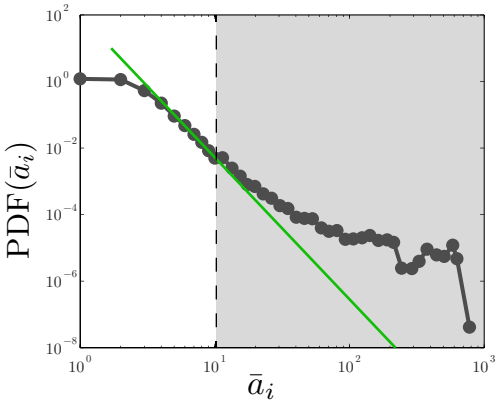


Figure 6.1: **Distribution of number of authors per paper.** For each paper i in the dataset, we denote with $\bar{a}_i$ the number of its authors and report the distribution $P(\bar{a}_i)$. The vertical line falls at ten authors, corresponding roughly to the point were the distribution deviates from the power law fitting line. We thus do not take into account in the disambiguation process all the papers that have more than 10 authors, corresponding to only 3% of all the papers.

are no more than 10 different authors. Criterion (i) is necessary to associate at least one affiliation to each author, which is a crucial step for the author name disambiguation; criterion (ii) is necessary to identify those publications where each author can be considered to have a substantial contribution which is crucial for our applications developed in chapters 7 and 8. The threshold of 10 authors has been chosen after the inspection of the distribution of number of authors per paper (Fig. 6.1). The probability density function can be approximated by a power-law, in line with previous studies [315]. We observe a deviation from the power-law for papers containing more than 10 authors, suggesting different retribution characterizing these large collaborations [316, 317]. The application of criteria (i) and (ii) gives us a final set of $425,369$ publications and $1,383,487$ occurrences of author names.

By applying the disambiguation procedure described above, we end up with a total of $237,038$ unique individuals, corresponding to a 83% decrease in the number of author name occurrences.

Accuracy

To evaluate the accuracy of our algorithm to disambiguate author names [315], we selected 200 pairs of papers that our algorithm predicted to

have been written by the same author, and 200 pairs for which the authors are predicted to be distinct individuals. We then determined how many times the pairs correspond to the same individual or not, by searching manually the authors homepage, scholar profile if any, looking at coauthors, affiliation, topic, etc. Out of these 400 pairs, we find the false positive rate (*i.e.* fraction of times the procedure indicates the pair of publications belonging to the same person, while they do not) to be 2% and a false negative rate (*i.e.* fraction of times that the same individual is considered to be two distinct persons) of 12%.

Complexity
The overall complexity of this algorithm is $O(n^2)$ where n is the number of author name occurences, i.e. number of papers multiplied by the average number of authors per paper. As only pairs of authors sharing the same last name and initials are compared to each other, that complexity is reached in the worst case when all authors in the dataset share the same last name and initials. However, in practice, this configuration is very unlikely as thousands of last names and first names are often present in the data and as a consequence the complexity is likely to be much smaller.

Semantic bias
The errors induced by disambiguation are not uniformly distributed among scientists. Indeed, as pointed out in previous research [307, 318–320], Asian names are the most difficult to disambiguate. Indeed, we examined 500 random pairs of papers from our subset and found that about 90% of manually detected errors are related to Asian names. This is not surprising as the 19 most common Chinese names, for example, represent about 56% of the population in China [303] or as 45% of Korean names are either *Kim*, *Lee* or *Park* [321]. Therefore, data related to asian authors should be taken with caution as many errors are undetectable.

6.3 Affiliation disambiguation

Exploiting geographical digital traces present in publication data is not only important to detect unique authors, it is also crucial to understand scientific mobility. While many studies explored the role of geography on scientific collaborations or mobility, most of them focused on collaborations or movements between countries or cities [299, 300] but not

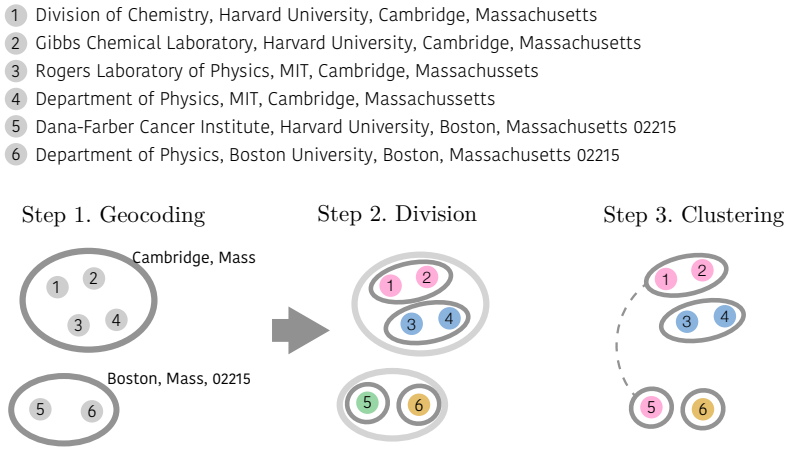


Figure 6.2: **Illustration of the affiliation disambiguation process** In the first step, affiliations are grouped by geographical locations. Affiliations within each group are then divided into subgroups based on their string similarity. Finally, subgroups of affiliations from different groups are iteratively merged based on their string similarity as well. At the end the process, three distinct institutions in our example are detected: $\{1, 2, 5\}, \{3, 4\}, \{6\}$

between institutions. However, studying movements in between research institutions would not only provide information at a much finer scale, but would also offer a better way to understand the influence of specific institutions on careers as well as their organisation. These aspects can hardly be explored with data characterised by a resolution at a country or city scale. Indeed, the Boston metropolitan area only, for example, contains 58 higher educational institutions, 9 of them being major universities with large ranking differences [322]. The effect of one of these institutions on individuals over another is thus undetectable at a city scale. Given the growing interest of studying career trajectories and collaborations at a finer scale we here propose a method that can efficiently disambiguate affiliations in publication data.

Method

In this method, we consider the collection of all distinct affiliation names occurring on all papers in the data. We consider the affiliation name as a string where the different fields, if any, are separated by a comma as it often appears on scientific articles and as illustrated on Figure 6.2. The process to disambiguate these names is divided into three agglomerative and divisive steps.

1. *Geocoding*: For this agglomerative step, we first geocode all affiliation names present in the data using the Google Geocoding API on the last two fields of their names. Each affiliation is thus associated to a set of coordinates that uniquely defines a location. We then group affiliations that share the exact same coordinates together (Fig.6.2). This first step usually dramatically decreases the number of comparisons between elements required for the next steps.

2. *Division*: In this second step, the basic idea is to compute a similarity measure between all pairs of affiliation names within a group. These similarity values are then used to divide each group into subgroups containing similar affiliations (Fig.6.2).

   In this method, we consider each affiliation name as a string. Hence, to compare affiliations, one has to design a method that returns the similarity between a pair of strings. In this step, we use the *cosine similarity* which is a vector-based measure of the similarity of two strings. The idea behind this measure is to transform each string into a vector in some high dimensional space such that similar strings are close to each other. The cosine of the angle between two vectors is a measure of how "similar" they are, which in turn, is a measure of the similarity of these strings.

   To transform each of our string, i.e. an affiliation name, into a vector, we use the popular *TF-IDF* vector representation. The *TF-IDF* vector is composed of the product of the *term frequency* (TF) and the *inverse document frequency* (IDF) for each word that appears in the string. The length of the *TF-IDF* vector is equal to the total number of unique words appearing in the corpus of affiliation names and the vector stores the *TF-IDF* value corresponding to each word for each string. The *term frequency* is the number of times the word appears in the string and is a measure of the importance of that word in the string. The inverse document frequency correspond to the inverse of the number of strings in which the word appears and serves to normalise the effect of words that appear commonly in many strings (such as "the" or "of"). The product of TF and IDF is thus a measure of the importance of the word in the string and the corpus as a whole. As affiliation names correspond to short strings compared to the total number of distinct words in the corpus, these vectors tend to be very sparse.

   For each pair of affiliation names within a group, we thus first

translate their two corresponding strings into their respective *TF-IDF* vectors and then normalise them into unit length vectors. The similarity between the two strings is then obtained by computing the cosine angle between the two corresponding vectors, which corresponds simply to the dot product as they are normalised.

In this particular step, we compute the *cosine similarity* between all pairs of affiliations within a group. Each time the similarity is larger than a particular threshold, denoted as κ_1, the two affiliations are merged together. At the end of the process, each group is thus divided into subgroups of similar affiliations (Fig 6.2).

The threshold κ_1 is determined by the following procedure. We randomly select two lists of 200 affiliation pairs: (i) pairs of affiliation within a group that are known to correspond to a single institution (manual check) and (ii) pairs of affiliations within a group that are known to correspond to two distinct institutions. The value of κ_1 is then given by the value of the threshold that minimises both the false positive rate (proportion of pairs from (ii) for which the similarity is over κ_1) and the false negative rate (proportion of pairs from (i) for which the similarity is under κ_1). This procedure ensures that the chosen threshold κ_1 leads to well disambiguated subgroups within each group.

3. *Clustering*: Departments or sub-units of a single institution are often not located at the same address, as illustrated in Figure 6.2 for Harvard University. As each subgroup from step 2 contains affiliations that share the exact same coordinates (due to the *geocoding* step 1), it is thus important to compare subgroups that belong to different groups, i.e. subgroups that do not share the same coordinates, in order to resolve this location issue (Fig.6.2).

   Hence, in this last step, pairs of subgroups belonging to different groups are compared to each other and merged accordingly by using a similarity value threshold κ_2 but also by using the author names previously disambiguated (see section 6.2). As an affiliation is associated to a particular author on a paper, for each scientist we compare his affiliations to each other and merge the two corresponding subgroups into one if the similarity is over the threshold κ_2. This particular approach is a key in our procedure. Indeed, similar affiliations on different papers authored by a same scientist are likely to correspond to a same institution. Moreover, this speeds up the algorithm by reducing the number of comparisons as only affiliation pairs associated to a same individual need to be

considered instead of all possible pairs. As mentioned earlier, this step is also important to reconnect institutions that are spread over different locations (e.g. Harvard University departments located in Cambridge and Boston as illustrated in Fig. 6.2).

The procedure to determine the optimal threshold κ_2 is identical to the one we use for the threshold κ_1 except that pairs of affiliations for the two lists (i) and (ii) are randomly taken from different groups and not within the same anymore.

Complexity

The overall complexity of our approach is $O(n^2)$ where n is the total number of distinct affiliation names present in the data. This complexity is reached in the worst case where all names share the same coordinates as n^2 comparisons would be required in the second step. However, the complexity is drastically lower in practice as most affiliations do not share the same coordinates.

Results on the American Physical Society dataset

A total of $319,829$ different affiliation names are identified in the APS dataset. As an example, $1,655$ of them are associated to MIT, which illustrates the difficult challenge associated to this process. This wide range of names results from different department and sub-department names within institutes but also historical changes (e.g. USSR), abbreviations and many typographical errors. Despite these inconsistencies inherent to this dataset, only $29,723$ geo-tagged groups are extracted in the first step of the algorithm, producing $64,107$ subgroups in the second step ($\kappa_1 = 0.9$). In the third step, an optimal value of $\kappa_2 = 0.7$ is found and results in a set of $4,052$ distinct affiliations.

Accuracy

To validate our results, we use a similar procedure than the one presented in section 6.2 for author names. We randomly select two lists of affiliation pairs: (i) 200 pairs of affiliations that are considered as a single institution and (ii) 200 pairs of affiliations located in the same city but considered as different by the algorithm. We then perform a search using publicly available information to determine and check, for each pair of affiliations, if they indeed correspond to similar institutions or not. Overall, the method exhibits a good accuracy. We find the false positive rate to be 11% (i.e. affiliations that are considered as a single institution while in reality they are not) and a false negative rate of 6% (i.e. affiliations that are wrongly categorised as distinct institutions).

6.4 Topic detection

As mentioned earlier, research on information networks often requires topic disambiguated data [299, 300]. In this section we address this issue by introducing a method that can efficiently detect scientific papers belonging to a similar scientific area or similar topic. By starting from a small core of disambiguated papers, we show how one can detect the corresponding community by taking advantage of the citation network over time. To prove its usefulness and accuracy, we then apply our approach on a set of 50 millions papers to detect those belonging to one particular scientific field: *Physics*.

Method
We consider a citation network constructed from publication data where a node $n_i \in N$ corresponds to an individual publication and where a link $(i, j) \in K$ exists if the publication corresponding to n_i contains a reference towards the publication corresponding the n_j. Each node n_i is thus characterised by an in-degree k_i^{IN} (number of citations) and an out-degree k_i^{OUT} (number of references). Nodes with $(k_i^{IN}, k_i^{OUT}) = (0, 0)$ are isolated nodes and are not considered in the network. Finally, each node n_i is characterised by a variable t_i corresponding to the time of publication of the article associated to the node.

The method is an iterative process where at each step s three sets of nodes are computed: C_s, T_s and E_s (Figure 6.3).

The first type of set, C_s, includes the nodes that are considered to be part of the target community at a given time step s by the algorithm. Initially, we consider $C_0 \subseteq N$, denoted as the initial *core* set, to contain nodes that are originally known to be part of the community (or topic) of interest. The purpose of this initial *core* set is to act as a seed to detect other nodes that are part of the community and that will be iteratively included in C_s at subsequent steps $s > 0$.

The second type of set, T_s, is denoted as the *tangent* set and is defined as follow

$$T_s = \{n_i | n_i \notin C_s \wedge ((i, j) \in K \vee (j, i) \in K) \wedge n_j \in C_s\}. \tag{6.1}$$

It contains all the nodes outside the core set C_s that have at least one connection from or to a node within C_s (Figure 6.3, blue nodes). Initially, T_0 is thus directly derived from C_0. The purpose of the set T_s is to contain all candidate nodes, i.e. nodes that might subsequently

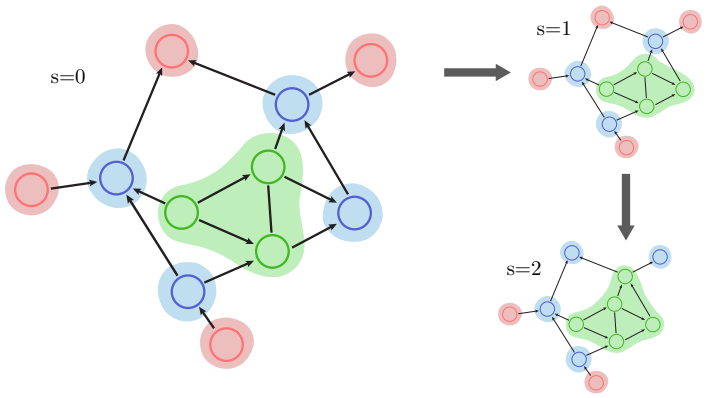


Figure 6.3: **Illustration of the topic detection method.** In this illustrative citation network, nodes within the core C_s are represented in green, while nodes of the tangent set T_s and external sets E_s are represented in blue and red respectively at each step s. Initially, C_0 is defined as the three green nodes (s=0). Following Eqs. 6.1 and 6.2, T_0 (blue) and E_0 (red) are computed. For the first iteration (s=1), the set C_0 is updated to C_1 by adding the nodes fulfilling the conditions defined by 6.9 and 6.10, resulting in updated sets T_1 and E_1. The same process is then repeated in the second iteration producing the updated sets C_2, T_2, E_2 (s=2). The process stops when no more nodes can be added to the core set C_s.

be added to the target community at step s after inspection of their incoming and outgoing links.

The third type of set E_s, denoted as the *external* set, is defined as

$$E_s = \{n_i | n_i \notin C_s \wedge (j,i) \notin K \wedge (i,j) \notin K \wedge n_j \in C_s\}. \tag{6.2}$$

and corresponds to nodes outside the core set C_s that share no connection with nodes within C_s (Figure 6.3, red nodes). Initially, E_0 is also directly derived from C_0. The basic idea behind the set E_s is to represent all non-candidates nodes, i.e. nodes that have no chance of being added to the community at step s. Note that by definition $C_s \cup T_s \cup E_s = N$ and $C_s \cap T_s \cap E_s = \emptyset$.

The basic idea of the method is to iteratively extend the target community C_s into C_{s+1} by adding candidate nodes from T_s that are expected to be part of the community. To do so, at each step s and for each node n_i we compute two variables: $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$. These variables

quantify the expectation of a particular node to be part of the target community (C_s) based on its incoming citations and outgoing references. The first variable $r_{i,s}^{IN}$, which focuses on incoming citations, is defined as

$$r_{i,s}^{IN} = \frac{k_{i,s}^{IN,\odot}}{\hat{k}_{i,s}^{IN,\odot}} \tag{6.3}$$

where $k_{i,s}^{IN,\odot}$ corresponds to the number of incoming citations to node n_i originating from nodes in C_s while $\hat{k}_{i,s}^{IN,\odot}$ corresponds to this *expected* number if links were randomly shuffled in the network. This expected number of incoming links to node n_i from the set C_s is given by

$$\hat{k}_{i,s}^{IN,\odot} = k_i^{IN} \frac{\sum_{n_j \in C_s} k_j^{OUT}}{\sum_{n_j \in N} k_j^{OUT}} \tag{6.4}$$

where k_i^{IN} denotes the total number of incoming citations to node n_i, i.e. its in-degree, and the remaining term corresponds to the probability for a link in the network to originate from C_s. This probability is simply given by dividing the number of outgoing links in the network originating from nodes in C_s ($\sum_{n_j \in C_s} k_j^{OUT}$) by the total number of outgoing links in the network ($\sum_{n_j \in N} k_j^{OUT}$). However, as a publication cannot be cited by older ones, only nodes n_j associated to publications published after n_i ($t_j > t_i$) are to be taken into account to compute this probability. Thus, Eq. 6.5 becomes

$$\hat{k}_{i,s}^{IN,\odot} = k_i^{IN} \frac{\sum_{n_j \in C_s | t_j > t_i} k_j^{OUT}}{\sum_{n_j \in N | t_j > t_i} k_j^{OUT}} \tag{6.5}$$

Similarly for outgoing references, $r_{i,s}^{OUT}$ is defined as

$$r_{i,s}^{OUT} = \frac{k_{i,s}^{OUT,\odot}}{\hat{k}_{i,s}^{OUT,\odot}} \tag{6.6}$$

where $k_{i,s}^{OUT,\odot}$ corresponds to the number of outgoing references from node n_i to nodes in C_s while $\hat{k}_{i,s}^{OUT,\odot}$ corresponds to this *expected* number in a random case. This expected number of outgoing links from n_i to the set C_s is given by

$$\hat{k}_{i,s}^{OUT,\odot} = k_i^{OUT} \frac{|C_s|}{|N|} \tag{6.7}$$

where k_i^{OUT} denotes the total number of outgoing references from node n_i, i.e. its out-degree, and the remaining term corresponds to the probability to reference a node that belongs to C_s. As in a random case each node as an equal chance of being referenced by another node, this probability is simply given by dividing the number of nodes present in the set C_s ($|C_s|$) by the total number of nodes in the network ($|N|$). Again, as a publication can only reference older ones, only nodes n_j associated to publications published before n_i ($t_j < t_i$) are to be taken into account to compute this probability. Thus, Eq. 6.5 becomes

$$\hat{k}_{i,s}^{OUT,\odot} = k_i^{OUT} \frac{|C_{s,t_j<t_i}|}{|N_{t_j<t_i}|}. \tag{6.8}$$

Taken together, $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$ offer two distinct values at each step s to evaluate the likeliness of a node n_i to be cited from or to reference the core set C_s. While a value of $r_{i,s}^{IN} = 1$ or $r_{i,s}^{OUT} = 1$ would indicate, respectively, that the number of incoming citations or outgoing references to the core set is not different than what we would observe in a random case, a value $r_{i,s}^{IN} > 1$ or $r_{i,s}^{OUT} > 1$ would correspond to a node that is more likely to reference/be cited from nodes from the core than what would be expected. Conversely, $r_{i,s}^{IN} < 1$ or $r_{i,s}^{OUT} < 1$ corresponds to a node that is less likely to reference/be cited from nodes from the core than what would be expected.

As a result, at each step s of the process, we use the variables $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$ associated to nodes in T_s to produce the updated core set C_{s+1}. Initially, C_{s+1} contains all the nodes from C_s. Then, for each node $n_i \in T_s$, we add n_i to C_{s+1} if one of the two following conditions is fulfilled

$$r_{i,s}^{IN} > \tau_1 \tag{6.9}$$

$$r_{i,s}^{OUT} > \tau_2. \tag{6.10}$$

The thresholds τ_1 and τ_2 are assigned based on a parameter p. Given p, the thresholds τ_1 and τ_2 correspond respectively to the p^{th} percentile of the distribution of $r_{i,0}^{IN}$ and $r_{i,0}^{OUT}$ values for nodes within the initial core set C_0 (Fig. 6.4). We choose the two distributions $r_{i,0}^{IN}$ and $r_{i,0}^{OUT}$ as a baseline for two main reasons: (i) we want to gauge the likeliness of candidates with the likeliness of *true* members, i.e. nodes for which we are sure they are part of the target community, simplifying the interpretation of the thresholds τ_1 and τ_2, (ii) as C_0 is fixed from the beginning, the thresholds will remain constant through the different iterations, reducing the complexity of the process.

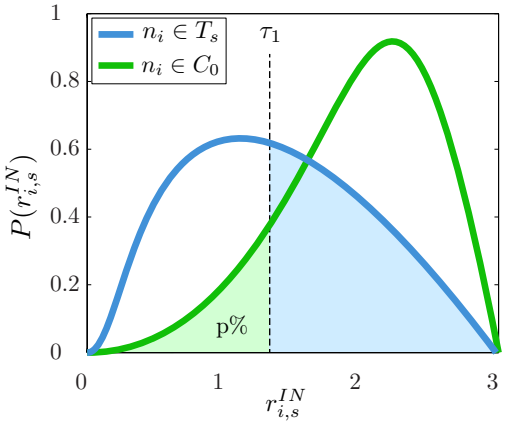


Figure 6.4: **Selection process of nodes from T_s added to the updated core set C_{s+1}.** An illustrative example of the distribution of $r_{i,s}^{IN}$ for nodes $n_i \in T_s$ (blue curve) and $r_{j,0}^{IN}$ for nodes $n_j \in C_0$ (green curve) is represented. All nodes $n_i \in T_s$ with a value $r_{i,s}^{IN}$ larger than the p^{th} percentile (τ_1) of the distribution of $r_{j,0}^{IN}$ for $n_j \in C_0$ are added to the updated core C_{s+1}. The added nodes are represented by the blue area, while the green area corresponds to nodes within the initial core set with a value $r_{i,0}^{IN}$ below the p^{th} percentile of their distribution. The same selection process is again applied to nodes but with distributions of $r_{i,s}^{OUT}$ and $r_{i,0}^{OUT}$

Being a percentile, the parameter p varies between 0 and 100% and can be considered as a *tolerance* parameter in the sense that it defines the minimal attraction needed for a node to be incorporated in the growing core. Indeed, a high value of p limits the addition of nodes $n_i \in T_s$ into C_{s+1} unless they have a sufficiently high value of $r_{i,s}^{IN}$ or $r_{i,s}^{OUT}$, while a low value of p would allow them to be added to the core (Fig. 6.4).

Once all nodes $n_i \in T_s$ satisfying the conditions (6.9) or (6.10) are added to the core set C_{s+1}, both sets T_s and E_s can be updated to T_{s+1} and E_{s+1} from C_{s+1} using Eqs. 6.1 and 6.2 and the next iteration can start. The process stops when C_s has converged, i.e. when no nodes from T_s can be added to the core set C_s. Note that while the thresholds τ_1 and τ_2 remain constant during the whole process, the values $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$ associated to each node n_i will change at each iteration, given their definition (Eqs. 6.3 and 6.6) and the fact that new nodes will incorporate the set C_s at each iteration step s.

Complexity
The overall complexity of the algorithm is $O(nm)$ where n denotes the number of nodes in the network and m the number of iterations. In the worst case, we have $m = n$ which would correspond to the situation where only one node from the *tangent* set T_s is added to the core set C_s at each iteration. However, in practice, we expect a much lower number of iterations especially if the citation network is characterised by a community structure.

Results on the Web of Science™ dataset
In order to illustrate our approach, we apply our topic-detection algorithm on a citation network constructed from the publication data of Web of Science™ to detect the subset of papers that pertain to physics. We restrict our approach to documents categorised as *article* in the dataset as other types of documents such as *editorial* or *abstract* do not always contain references. This leads us to a citation network containing 25 million articles and about 400 million citations.

Following the process of the algorithm, we start by identifying nodes that will form the initial core set C_0. Since we are interested in detecting the physics community of articles, we have to add nodes corresponding to articles that are known to be labeled as Physics. To do so, we identify 295 scientific journals present in our dataset that are considered as physics journals. To identify these, we first extract the list of scientific journals categorised as *Physics* on four major websites: Elsevier [323], Springer [324], Wikipedia [325] and PhysNet [326]. We then manually check each journal to ensure that they actually contain only physics articles. All nodes associated to articles published in one of these journals, amounting to 2,438,284 nodes, represent the set C_0. Following Eqs. 6.1 and 6.2, we build the two other initial node sets T_0 and E_0. The set T_0, which contains nodes that are outside C_0 and that have at least one citation from or a reference to a node within C_0, is represented by 4,296,286 nodes. On the other hand, the number of nodes outside C_0 that are not connected to nodes within C_0 amounts to 18,824,194 nodes and represent E_0.

In order to understand the role of the parameter p in our approach, we apply the algorithm with eight distinct values of p (2, 5, 10, 20, 40, 60, 80 and 100). Given the initial node sets, C_0, T_0 and E_0, we iteratively compute the updated node sets C_s, T_s and E_s for each value of p until C_s converges. As expected, the smaller p is, the larger the core set C_s

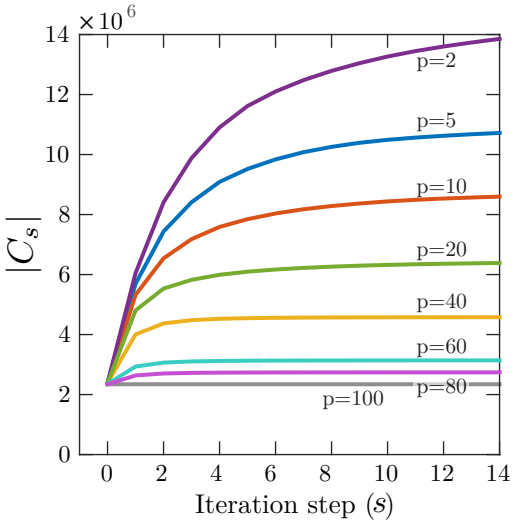


Figure 6.5: **Growth of core set C_s for different parameter p.** Size evolution of the core set C_s at different iteration steps of the process and for different parameter p. Smaller values of p lead to larger core sets as thresholds τ_1 and τ_2 are lower and more nodes can be added to the core (Fig. 6.4)

gets (Figure 6.5). This growth effect is due to the conditions (6.9) and (6.10) as smaller values of the parameter p give smaller thresholds τ_1 and τ_2, resulting in a less restrictive definition of the community.

To assess the accuracy of our method, we select articles published in two major interdisciplinary journals: Science (1995-2013) and PNAS (1915-2013). We then divide these articles into two different sets. The first set corresponds to 2,062 articles categorised in the *Physics* section of both journals, while the second set corresponds to 3,715 articles classified in an other section (e.g. medicine, genetics, economics, etc) in both journals. For each set, we compute the proportion of papers that are part of the final core set C_s for each value of the parameter p (Fig. 6.6). For most values of p, we see a remarkable accuracy as almost no non-physics articles are present in the detected community. Moreover, the observed increase of non-physics papers for $p = \{2, 5\}$ is not surprising as these correspond to articles that were originally classified in a different section than *Physics* but which later emerged as interdisciplinary physics articles (e.g. Economic paper which became relevant to Network Science, Table 6.1.A). Regarding the set of articles classified in the *Physics* section, we observe the same phenomena as articles are

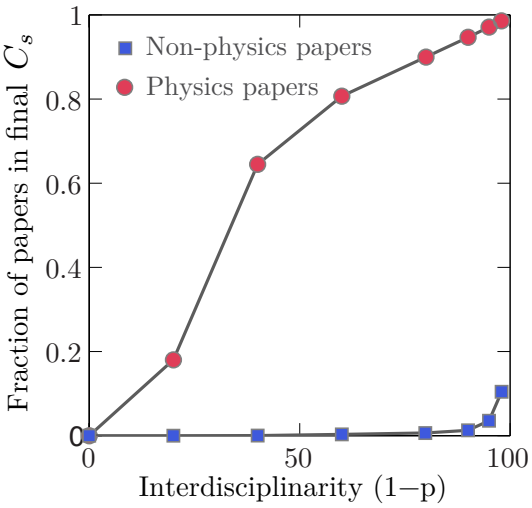


Figure 6.6: **Accuracy to detect physics vs non-physics articles.** Two sets of articles from interdisciplinary journals are constructed: (i) 2,062 articles classified as *Physics* and (ii) 3,715 articles categorised in an other section. The proportion of articles from (ii) is remarkably low for different values of p (blue points), demonstrating the excellent accuracy of the algorithm. On the other hand, articles from (i) are gradually incorporated according to their degree of interdisciplinarity. High values of p include articles that are considered as "pure" physics, while lower values of p incorporates interdisciplinary articles as well.

Table 6.1: **Example of scientific articles**

A	*Zipf distribution of US firm sizes, Science*, 2001
B	*On the Green's functions of quantized fields*, PNAS, 1951
C	Two-dimensional atomic crystals, PNAS, 2005
D	Hyper telomere recombination accelerates replicative senescence and may promote premature aging, PNAS, 2008

gradually incorporated according to their degree of interdisciplinarity. High values of p include articles that are considered as "pure" physics (e.g. Field theory, Table 6.1.B), well referencing/cited by articles within the community. On the other hand, lower values of p incorporate interdisciplinary articles as well (e.g. Graphene research, Table 6.1.C), less referencing/cited by articles within the community. Again, this exhibits the excellent accuracy of the approach as 99% of the articles classified in the *Physics* section of *Science* and *PNAS* are included in C_s for $p = 2$. A more detailed search reveals that the few missed physics articles mostly correspond to either articles published before 1930 where few statistics (citations and references) are available or either articles that later emerged as being important in other scientific fields (e.g. Genetics, see Table 6.1.D). As a result, our variable p can be seen as a parameter that sets the interdisciplinarity tolerance of the target community. This characteristic is particularly important as the delimitation of these types of communities, i.e. topics, is subjective and articles can often belong to more than one particular field or topic.

6.5 Conclusion

In summary, we presented three distinct yet related approaches to deal with data ambiguity. We first introduced an agglomerative algorithm that can efficiently disambiguate author names in very large publication data. We validated this method on real publication data extracted from the American Physical Society, obtaining accurate disambiguated groups of scientists. Developing disambiguation techniques for names is extremely important at the present time as the field of *Science of Science* is rapidly developing and more and more publication datasets become available to researchers. As evidence of the method's utility, the disambiguated author names obtained in this chapter are used as a baseline in the studies developed in Chapters 7 and 8.

In the second part of this chapter, we addressed the issue of ambiguities in location names by describing an agglomerative and geo-tagged approach to accurately detect similar affiliations in publication data. This issue is crucial for studies on scientific mobility but not only. As the use of social and location-based mobile and web applications are becoming part of our everyday routine, generating more and more mobility data, the need for tools to manage ambiguities and classify locations in these large data is quickly growing, both in academia and in the industry. With its flexibility as well as its simplicity of implementation, our approach can offer such a tool to researchers and actors in the data analytics industry.

Finally, we introduced a network-based algorithm to detect a community of articles related to the same scientific field. This algorithm shares similarities with "label propagation" methods which are common in Machine Learning [327]. We implemented our approach on a real-world citation network characterised by more than 25 million nodes and 400 million citation links to detect the *Physics* community of papers. Not only the algorithm accurately uncovers correct articles but it also offers a way to control for the interdisciplinarity tolerance of the community. Beyond the accuracy of the results, the variables developed in the algorithm offer novel measures to understand the role and evolution of particular articles within the community. For example, we can easily detect articles that were not originally meant to be about physics (low $r_{i,s}^{OUT}$ value) but which later came out to be relevant to the community or which initiated a new scientific branches within physics (high $r_{i,s}^{IN}$). These measures coupled with citation information offers tremendous opportunities to understand the evolution as well as the profound changes that emerged in physics but also in any other fields. Even though we focused on publication data in this chapter, this method can also be applied to other types of networks and can offer valuable insights for decision-makers in marketing or even politics.

Chapter 7

Quantifying patterns of scientific success

In this chapter, we study the quantitative patterns of scientific performance through the analysis of two distinct aspects of an individual's career: his productivity and impact. First, we define how scientific impact is defined, allowing us to detect the highest impact work of individuals. Second, we quantify the changes of productivity throughout a scientist's career, showing that science is not different than other areas of human performance. Finally, we analyse the changes of impact in scientific careers, finding that impact is distributed randomly within a scientist's sequence of publications.

7.1 Introduction

Motivation The path to major accomplishments in most areas of human performance, from sport to music, poetry or engineering, usually requires a steep learning curve, long practice and many trials [47]. Athletes go through demanding training and participate in many competitions before setting new records; musicians practice since early age and perform in secondary venues before earning the spotlight [48]; programmers participate in numerous routine tasks before addressing more innovative projects. This gradual increase in performance through learning and practice characterizes most innovative trades [328] and common sense suggests this to be true in science as well: the outstanding discoveries a scientist is known for are typically preceded by results and papers of less memorable impact. Indeed, despite the many discoveries they may have made during their careers, scientists tend to be remembered for a single discovery: their highest impact work. We know Alexander

Fleming for his discovery of penicillin [329], Marie Curie for her research on radioactivity [330], Emmy Noether for the connection between symmetry and conservation laws [331], or Crick and Watson for the double helix [332]. This prompts us to ask: What are the precise patterns that lead to scientific success? Does performance indeed improve throughout a scientific career? Are there quantifiable signs of an impending scientific hit? Will a scientist, having made a major discovery, produce higher impact paper than before his/her breakthrough?

In this chapter, we explore these questions by quantifying the changes in productivity and impact induced by a scientist's highest impact work. First, using citation-based measures as a proxy of impact [46, 333–336], we show how we identify the highest impact paper in each scientist's career. We then show that, while there are reproducible productivity patterns leading to the highest impact work, surprisingly major discoveries are not preceded by works of increasing impact, nor are followed by work of higher impact. Yet, we observe that the emergence of the highest impact paper is not entirely random either: while the sequence of papers' impact in each individual career appears to be dramatically unpredictable, scientists with truly outstanding publications [335] have higher productivity [49, 337], and a different paper impact distribution, indicating that there are statistical features peculiar to outliers. Finally, we conclude this chapter by discussing the obtained results and suggesting future potential research investigations.

Data We consider papers published by the American Physical Society (APS) from 1893 to 2010 (see section 6.1 for details), a record containing 425,369 publications in 11 journals. We used the disambiguation technique described in section 6.2 to infer author identity, compiling the accurate publication history of 237,038 scientists (Fig. 7.1A). To eliminate authors that abandon research at an early stage of their career and to have enough statistics for each individual, in this chapter we limit our analysis to scientists that (i) have authored at least one paper every 5 years, (ii) have published at least 10 papers, (iii) their publication career spans at least 20 years [338, 339], arriving to 2,887 scientists with persistent publication record.

7.2 Measure of paper impact

Citation-based measures of impact are affected by two major problems: (1) citations follow different dynamics for different papers [55, 333] and (2) the average number of citations changes over time [340]. To overcome (1) for each paper we use the cumulative number of citations the paper received 10 years after its publication, c_{10}, as a measure of its scientific impact [55, 333, 341]. We can correct for (2) by normalizing c_{10} by the average c_{10} of papers published in the same year, but this correction does not alter our conclusions, hence we report in this chapter results without normalization. To calculate c_{10} we limit the study to publications published up to the year 2000. This requisite together with the career span limit of 20 years implies that the studied scientists started their career in 1980 or before.

Highest impact paper
To capture potentially outstanding impact, for each researcher we identify his/her most cited paper, c_{10}^*, or the paper with the highest number of citations ten years after its publication. We denote with t^* the time of publication of this work and with N^* its position in the sequence of N papers published by the scientist during his/her career (Fig. 7.1A). The distribution $P(c_{10}^*)$ for all scientists indicates that only 5% have $c_{10}^* \geq 200$, hence most careers are characterised by limited peak impact. To systematically distinguish the careers based on their peak impact, we group each scientist into high impact (top 5%, $c_{10}^* \geq 200$), low impact (bottom 20%, $c_{10}^* \leq 20$), and normal impact (middle 75%, $20 < c_{10}^* < 200$) categories (Fig. 7.1B).

7.3 Patterns of productivity

Cumulative productivity, or the total number of papers a scientist i publishes up to time t after his/her first publication, is known to asymptotically follow (Fig. 7.2) [49]

$$N_i(t) \sim t^{\gamma_i}. \tag{7.1}$$

We find that for low impact scientists $\langle\gamma\rangle = 1.55$, indicating a steady increase in their productivity. The increase is much faster for high impact researchers, however, for whom $\langle\gamma\rangle = 2.05$ (Fig. 7.3A). These trends

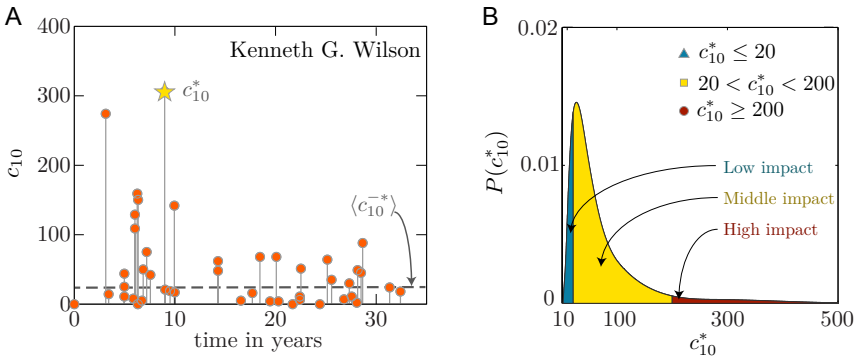


Figure 7.1: **Definition and distribution of highest impact papers** **(A)** Publication history of Kenneth G. Wilson (Nobel Prize in Physics, 1982). The horizontal axis indicates the number of years after the scientist's first publication and each vertical line corresponds to a research paper. The height of each line corresponds to c_{10}, *i.e.* the number of citations the paper received after 10 years. The highest impact paper of Kenneth Wilson was published in 1974, 9 years after his first publication and it is 17th of his 48 papers, hence $t^* = 9$, $N^* = 17$, $N = 48$. **(B)** Distribution of the highest impact paper $P(c_{10}^*)$ across all scientists, fitted by a lognormal function (continuous line). We highlight in blue the bottom 20% of the area, corresponding to low impact scientists ($c_{10}^* \leq 20$); the red area indicates the high impact scientists (top 5% , $c_{10}^* \geq 200$); yellow corresponds to the remaining 75% middle impact scientists ($20 \leq c_{10}^* \leq 200$).

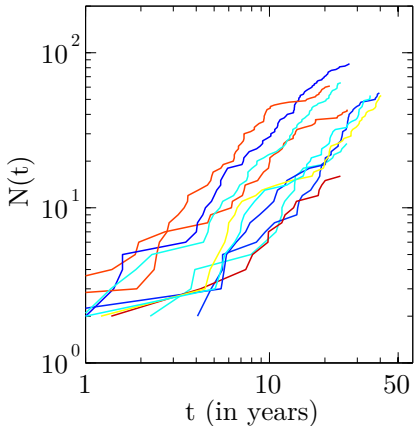


Figure 7.2: **The number of papers $N(t)$ up to time t for 10 randomly chosen scientists, capturing different productivity increase.** The paper publication date is known with a time resolution of days; $t = 0$ coincides with the day of the scientist's first published paper. Each curve can be asymptotically fitted with $N(t) \sim t^{\gamma}$, as indicated in Eq. 7.1 [49]. For each scientist, we extract the exponent γ based on the cumulative productivity in the second half of his/her career paper sequence.

are confirmed by the yearly productivity $\langle n(t) \rangle$: for high impact scientists productivity increases almost threefold during their career, while the increase is modest for low impact scientist (Fig. 7.3B). If, however, we explore productivity $\langle n(t) \rangle$ in the vicinity of year t^*, when a scientist publishes his/her most cited work c_{10}^*, we find that $\langle n(t) \rangle$ surges before t^* and drops following it, suggesting that the productivity improves as a scientist nears his/her highest impact work and drops afterwards. In other words, the high impact work appears to be a singular event in a scientist's career, influencing productivity (Fig. 7.3C) [293, 337].

Taken together, in line with previous work on productivity [47], Figure 7.3A-C confirms that productivity improves throughout a scientific career. We find, however, that this trend is modulated by impact: The productivity growth is particularly pronounced for high impact scientists and much weaker for low-impact scientists (Fig. 7.3A-B), and productivity appears to peak in the vicinity of the most cited work (Fig. 7.3C). Yet, long-term scientific excellence is rarely measured by productivity alone, prompting us to explore the similar patterns that describe the evolution of impact.

7.4 Patterns of impact

As illustrated in Fig. 7.1A, the scientists' careers are represented as time series, where each data point corresponds to a publication and the intensity is characterized by its impact. To detect impact trends before and after the c_{10}^* peak, two standard techniques can be applied to these time series [342].

The first technique is moving average, which provides a series of averages over different windows of the original time series. Given a series of numbers and a fixed window of size L, the first element of the moving average is obtained by taking the average of the initial L numbers in the series. Then the window is modified by "shifting it forward", that is, excluding the first number of the series and including the next number following the original subset in the series. This process is repeated over the entire time series, finally providing a new time series made of all averages, where short-term fluctuations are smoothed out. The second technique that can be used is a record series. Similar to the moving average, it produces a new time series by rolling a window L on the original data. The difference between this method and moving average

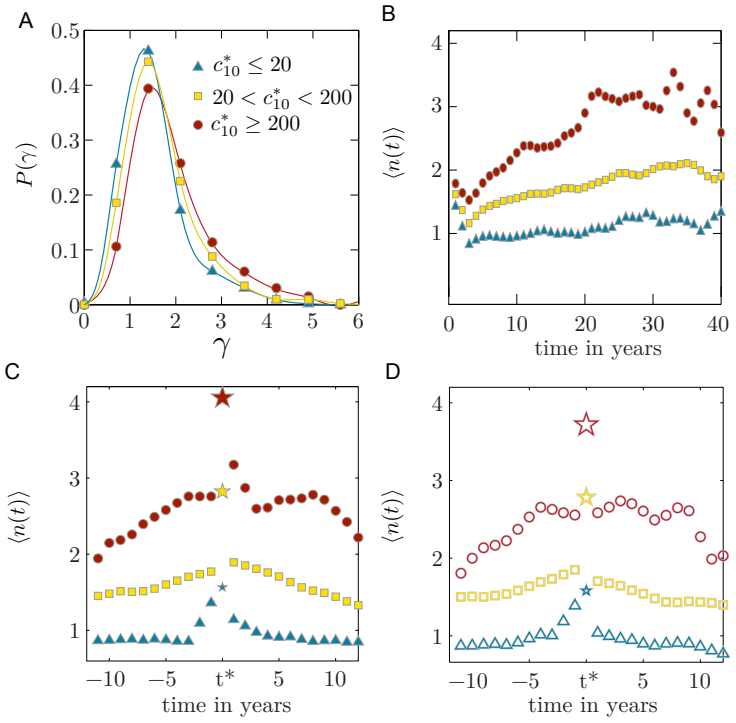


Figure 7.3: **Patterns of productivity during a scientific career.** **(A)** Distribution of the productivity exponents γ defined in (7.1) [49]. **(B)** Dynamics of productivity, as captured by the average number of papers $\langle n\,(t)\rangle$ published each year for high, middle and low impact scientists. $t = 0$ corresponds to the year of a scientist's first publication. **(C)** Dynamics of productivity captured by the average number of papers $\langle n\,(t)\rangle$ published the years before or after t^*, the publication year of the highest impact paper, for high, middle and low impact scientists. **(D)** Same as (C), but for a control dataset that randomly mixes the impact of each paper within a scientist's career.

is that the maximum value of the L numbers is considered in this case.

We apply both techniques using different values of L ($L = 1, 5, 20$) to all points but c^* of individual time series. For each resulting set of time series, we consider the average $\langle c_{10} \rangle$ before and after t^*, time of the highest impact paper, for individuals having similar c_{10}^* (Fig. 7.4). We observe a behavior that is robust against the choice of L: there is no discernible change of impact before nor after the publication of the highest impact work.

At the same time, if we look at yearly impact of scientists (Fig. 7.5), impact appears to follow similar patterns to productivity (Fig. 7.3B): the number of citations per paper increases during a high-impact scientist's career, an effect that is hardly noticeable for normal and low impact individuals. Yet, in Figure 7.4, we observe a dramatic deviation from productivity (fig. 7.3C) if we examine the impact in the vicinity of t^*, the publication time of the most cited work c_{10}^*. Indeed, we do not see a gradual increase in impact as a scientist approaches t^*, nor do we observe elevated citations following this breakthrough. Instead the observed pattern exhibits a singular behavior.

Finally, we randomize each career by leaving all productivity measures (N and $n(t)$) unchanged, but shuffling the impact of each paper within each career (Fig. 7.6). As we can observe, the papers published before and after t^* show no discernible differences in their average number of citations, but more importantly the lack of differences between the original and the randomized careers supports our overall conclusion: in contrast with productivity, there are no detectable changes in impact leading up to or following a scientist's highest impact work.

Timing of impact

To better understand the role of the peak impact on a career, we measure the probability $P(t^*)$ that the highest impact paper is published at time t^* after a scientist's first publication (Fig. 7.7A). The high $P(t^*)$ between 0 and 20 years indicates that most physicists publish their highest impact paper early or mid-career, in line with earlier findings for Nobel Laureates [343]. The drop in $P(t^*)$ after 20 years suggests that it is unlikely that a physicist's most cited work will come late in his/her career.

Yet, productivity is not uniform in time (Fig. 7.3B-C), prompting us to ask whether the peak in $P(t^*)$ may be rooted in changes in pro-

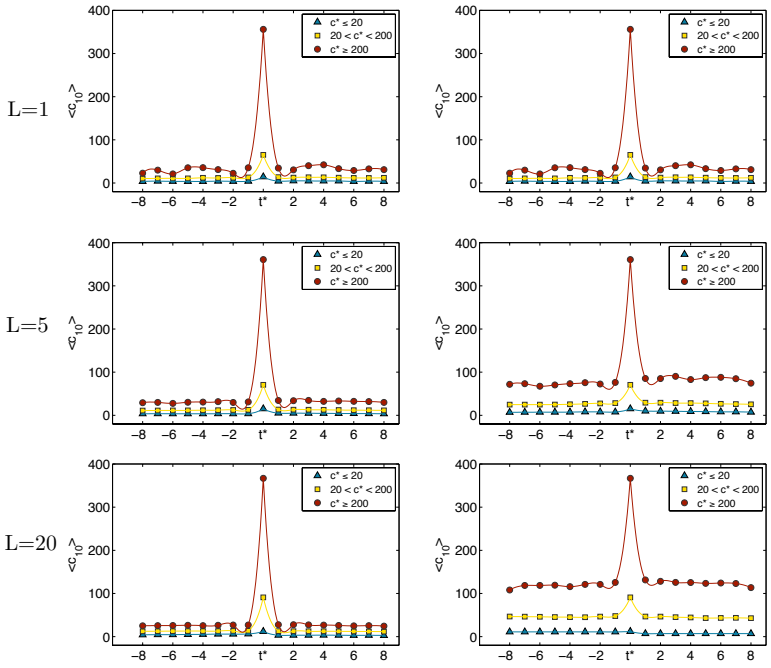


Figure 7.4: **Patterns of impact before and after publication of the highest impact paper.** We smooth the scientists' career by using a moving average (left panels) and a moving record (right panels), for different length of windows L. When $L = 1$ the careers are not smoothed and the finger curve is computed directly on the original data. For various values of L in both techniques no qualitative difference is observed for the different groups of scientists, as no patterns preceding and following the highest impact paper appear.

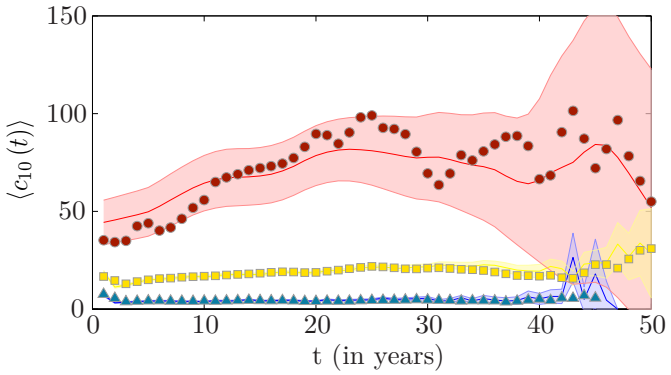


Figure 7.5: **Yearly impact of scientists.** The dynamics of impact captured by the yearly average impact of papers $\langle c_{10}(t)\rangle$ for high, middle and low impact scientists (see Fig. 7.1B for definition), where $t = 0$ corresponds to the year of a scientist's first publications. The symbols correspond to the data, while the shaded area indicates the 95% of confidence limit of careers where the impact of the publications is randomly permuted within each career.

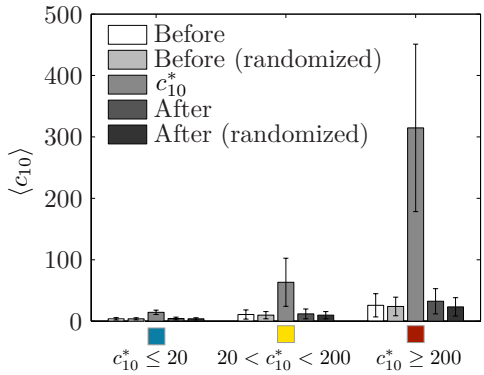


Figure 7.6: **Average impact before and after publication of the highest impact paper.** $\langle c^*_{10}\rangle$ and $\langle c_{10}\rangle$ before and after a scientist's most cited paper. For each group, we calculate the average impact of the most cited paper, $\langle c^*_{10}\rangle$ as well as the average impact of all papers before and after the most cited paper. We also report the same measures obtained in publication sequences for which the impact c^*_{10} is fixed, while the impact of all other papers is randomly permuted.

ductivity. Therefore we shuffled c_{10} among all papers published by the same scientist, preserving a scientist time-dependent productivity and paper-by-paper impact and randomizing only the order of his/her papers. The fact that $P(t^*)$ for these synthetic careers is indistinguishable from the original data (Fig. 7.7A) indicates that variations in $P\left(t^*\right)$ are fully explained by the higher productivity in the early stage of a career.

These results prompted us to explore $P\left(N^*/N\right)$, i.e. the probability that the most cited work is early (N^*/N small) or late ($N^*/N \simeq 1$) within the sequence of papers published by a scientist. We find that $P\left(N^*/N\right)$ is flat (Fig. 7.7B inset), a finding quantitatively supported by the cumulative $P(\geq N^*/N)$ (Fig. 7.7B), which decreases, independently of impact, as $(N^*/N)^{-1}$, fully in line with a uniform $P\left(N^*/N\right)$. Taken together we arrive at a rather unexpected conclusion, which we call the *random impact rule*, representing the main empirical finding of this chapter: impact is randomly distributed within a scientist's career, regardless of publication time or order in the sequence of publications.

The random impact rule prompted us to revisit our earlier finding that productivity peaks around t^* (Fig. 7.3C) and that impact of high impact individuals grows during their career (Fig. 7.5). We randomly shuffled again the impact of the papers within each career, while leaving the individual productivity unchanged and repeated the measurements of Figure 7.3C. The obtained productivity curve (Fig. 7.3D) is indistinguishable from the original data (Fig. 7.3C), indicating that there is no causal association between the timing of the highest impact paper and productivity. Similarly, the randomized careers display impact variations that are indistinguishable from the original data for both high and low impact individuals (Fig. 7.5 shaded-areas). Hence the growing impact of Figure 7.5A has a single explanation: as the productivity of the high impact individuals raises during their careers, their chance of drawing a high impact paper grows because of the increasing number of trials. As impact distribution is fat-tailed, $\langle c_{10} \rangle$ is not stable, but increases with the number of publications, resulting in the impact growth observed in Fig. 7.5A. Hence growing productivity, rather than changing impact, accounts for the observed career trends.

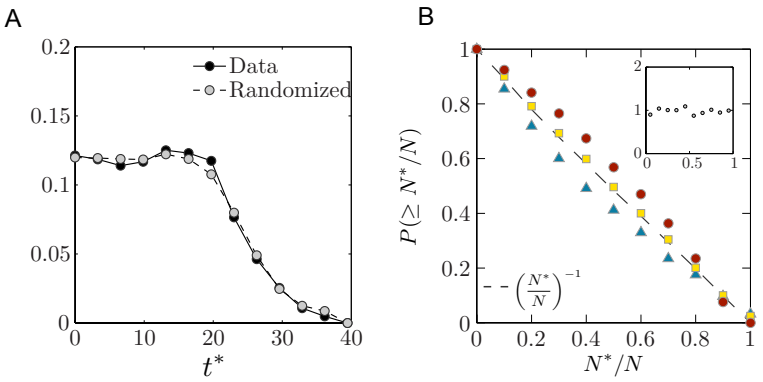


Figure 7.7: **Timing of highest impact work. (A)** Distribution of the publication time t^* of the highest impact paper for scientists' careers (black circles) and for randomised impact careers (grey circles). The lack of differences between the two curves ($p = 0.70$ for the Mann-Whitney U test between the two distributions) inspired the random impact rule. Note that the drop after 20 years is partly due to the the fact that we focus on careers that span at least 20 years. **(B)** Cumulative distribution $P(\geq N^*/N)$, where N^*/N denotes the order N^* of the highest impact paper in a scientist's career, varying between $1/N$ and 1. The cumulative distribution of N^*/N is a straight line with slope -1, indicating that N^* has the same probability to occur anywhere in the sequence of papers published by a scientist. The flatness of $P(N^*/N)$ (all scientists, inset) supports the conclusion that the timing of the highest impact paper is uniform.

7.5 Conclusion

In summary, the literature on innovative or high-performing careers suggest that performance should gradually improve during a scientist's career. Furthermore, following their highest impact work, scientists should maintain a career of elevated performance. Our measurements fully support these hypotheses if performance is measured in productivity: we find that productivity increases as a scientist approaches his/her highest impact work and it remains steady afterwards. There are also strong correlations with the magnitude of impact: the productivity boost is minimal for low impact careers, but it is remarkable for high impact scientist.

Yet, when we measure performance in terms of impact, as captured by citations, there are no discernible increase leading to the highest impact work, nor is there elevated impact following it. Rather, the highest impact work stands out as a singularity in the scientists' career. In fact the highest impact work can be with the same probability anywhere in the scientist's career. This random impact pattern, coupled with the increased productivity around 10 years after the beginning of a scientific career, is responsible for well-documented finding that the highest impact work emerges about a decade into a creative career [47]. This pattern can be fully explained by the random impact rule and an increased mid-career productivity.

As observed in Figure 7.5, there is a correlation between the magnitude of the highest impact work (c_{10}^*) and sustained performance: high impact scientists ($c_{10}^* \geq 200$) maintain a career of elevated performance. This raises several open yet important questions. How can sustained exceptional performance in terms of productivity or impact be explained ? Are there patterns of long-term success that remain to be determined? These two questions are legitimate as several aspects of the research environment can play a role on long-term success. However, our understanding of this role remains poor. In the next chapter, we address one aspect of this problem by analysing the mobility of scientists and quantifying the influence institutions can have on their performance.

Chapter 8

Impact of mobility on scientific success

Changing institutions is an integral part of an academic life. Yet little is known about the mobility patterns of scientists at an institutional level and how these career choices affect scientific outcome. In this chapter, we examine 450,000 papers to track the affiliation information of individual scientists, allowing us to reconstruct their career trajectories over decades. We first show that career movements are not only temporally and spatially localized, but also characterized by a high degree of stratification in institutional ranking. We then demonstrate that while going from elite to lower-rank institutions on average associates with modest decrease in scientific impact, transitioning into elite institutions does not result in subsequent impact gain.

8.1 Introduction

Despite their importance for education, scientific productivity, reward and hiring procedures, our quantitative understandings of how individuals make career moves and relocate to new institutions, and how such moves shape and affect performance, remains limited. Indeed, previous research on migration patterns of scientists [344, 345] tended to focus on large-scale surveys on country-level movements, revealing long-term cultural and economical priorities [346–349]. As highlighted in the first part of this thesis, research on human dynamics and mobility has emerged as an active line of enquiry [96, 97, 100, 114, 268, 289, 350], owing to new and increasingly available massive datasets providing time resolved individual trajectories [65]. While these studies cover a much shorter time scale than a typical career, they uncover a set of regularities and

reproducible patterns behind human movements [96–98]. Less is known about patterns behind career moves at an institutional level and how these moves affect individual performance.

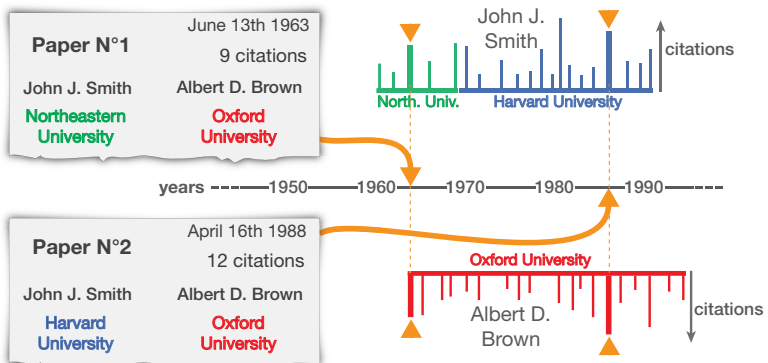
Here we take advantage of the fact that scientists publish somewhat regularly along their career [49, 50], and for each publication, the institution in which the work was performed is listed as affiliations in the paper, documenting career trajectories at a fine scale and in great detail. These digital traces, offering data on not only individual scientific output at each institution but also career moves from one institution to another, can provide insights for science policy, helping us understand how institutions shape knowledge, the typical moves of individual career development and help us evaluate scientific outcomes associated with professional mobility.

8.2 Resolving individual career trajectories

To extract individual career trajectories of scientists, we examine 450,000 publication data extracted from 11 journals of the American Physical Society (see section 6.1 for details about the dataset) and from which 212,316 unique scientists as well as 4,052 institutions are extracted with the algorithms introduced in section 6.2 and 6.3. To reconstruct the career trajectory of a scientist, we use the affiliation given in each of his/her publications (Fig 8.1). By analysing all publications from a particular author we are able to reconstruct his/her career trajectory. For authors with multiple affiliations listed on a paper we consider the first affiliation as primary institution. In order to detect career movements, i.e. changes in a scientist's institution, one has to remove artificial movements induced by short-term stays and by errors and typos in the affiliation names on the papers. To do so, only institutions reported in at least two consecutive papers are considered in a career trajectory. We compute the impact of each paper by counting its cumulative citations collected 5 years after its publication [55, 143, 292, 351].

8.3 Institutional performance

Over the past decades, we have witnessed a remarkable increase in systems that compare universities [352, 353]. Due to the growing acces-



	A_i	P_i	C_i
Northeastern University	+1	+1	+9
Harvard University	+1	+1	+12
Oxford University	+1	+2	+21

Figure 8.1: **Illustrative example of career trajectory reconstruction for hypothetical authors.** Given the paper *N°1* and *N°2*, we know that the author *John J. Smith* was affiliated to Northeastern University in 1963 and Harvard University in 1988. Extracting information from all his other publications allows us to reconstruct his career trajectory and discover that he was affiliated to Northeastern University for 8 years where he published 5 papers and then moved to Harvard University for 23 years where he published 16 papers. The cumulative number of citations of a paper obtained within 5 years after the publication is also known.

sibility to higher education institutions, the increasing demand for information on academic quality has led to the development of various international rankings, fueling international scientific mobility. Indeed, as the best lab for the type of research you are doing is usually not where you are [354–356] changing countries is now a rite of passage for many young researchers who follow the resources and facilities [49, 346]. Despite this growing interest for institutional rankings, many are skepticals about their validity and accuracy [352, 357, 358] as the dynamics behind institutional performance remains unclear. In the light of these events, we present in this section statistical results for institutions extracted from our disambiguated dataset. We first analyse the population size distribution as well as how citations are distributed among them. We then study the relationship between these two characteristics, highlighting the influence of population size on impact and productivity of institutions. We finally propose a ranking of institutions based on our findings which will be used in the next section to investigate the impact of institutions on individual career trajectories.

Three characteristics are computed for each institution i (Fig. 8.2): the institution size (A_i), representing the total number of distinct authors that published at least one paper at institution i; the number of papers (P_i) published under affiliation i; the cumulative number of citations (C_i) collected by all papers P_i. We find that $P(A)$ follows a fat tailed distribution, indicating significant population heterogeneity among different institutions (Fig. 8.2A). While most institutions are small, a few have a large number of scientist, often corresponding to large institutes or universities. We observe similar disparity in $P(C)$ (Fig. 8.2B): few institutions acquire a large number of citations, while most research labs or universities receive few citations.

Figures 8.3A-B show the correlation between the institution size A and both the average publication impact C/P and the average productivity P/A of institutions. The average productivity and impact of an institution are different but complementary measures of scientific performance. We find the institution size has little influence on productivity ($R^2 = 0.43$) (Fig. 8.3B), yet it positively correlates with the impact of publications ($R^2 = 0.85$), indicating that large institutions offer a more innovative/higher impact environment than smaller ones as captured by citations per paper (Fig. 8.3A). Also, as larger institutions develope more internal collaborations, the number of co-authors in publications from large institutions might be larger and, as a consequence, attracts more citations [351].

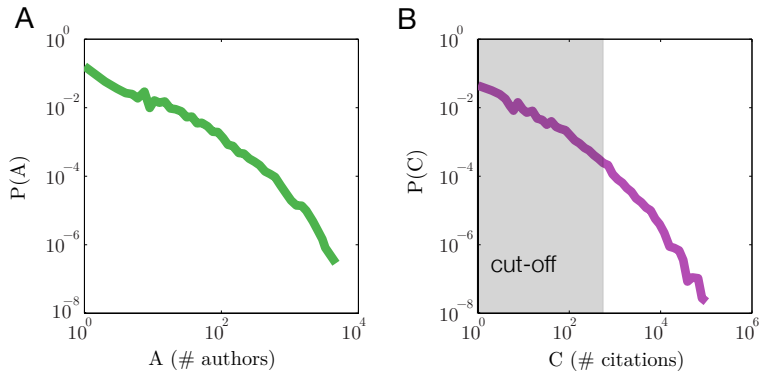


Figure 8.2: **Basic features of research institutions. (A)** The probability density function of institution size, A, follows a fat tailed distribution, indicating a significant heterogeneity. While most institutions size are small, a few have a large population, often representing large institutes or universities with a long history. **(B)** The probability density function of citations of institutions, C, is also very heterogeneous. Few institutions acquired a large number of citations, while most research labs or universities received few citations. Only the first thousand locations are taken into account in further analyses (shaded area)

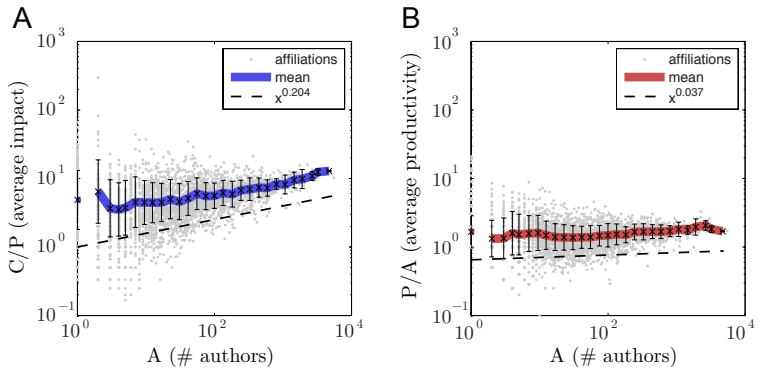


Figure 8.3: **Correlation between institution size, scientific impact and productivity (A)** The correlation between institution size and average publication impact is reported. Institution size positively correlates with the impact of publications ($R^2 = 0.9$) , indicating that large institutions offer a more innovative/higher impact environment than smaller ones as captured by citations per paper. The dashed line indicates a power-law behaviour with exponent $\alpha = 0.204 \pm 0.006$ **(B)** The correlation between institution size and institution average productivity is also reported, indicating institution size has little influence on productivity ($R^2 = 0.43$). The dashed line indicates a power-law behaviour with exponent $\alpha = 0.037 \pm 0.003$.

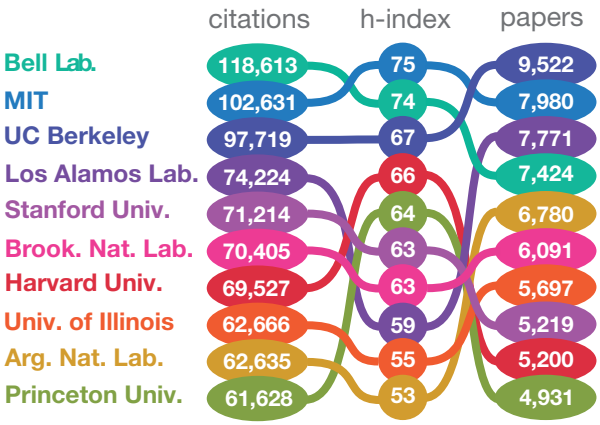


Figure 8.4: **Ten most cited institutions in physics.** Comparison between different rankings. The H-index is closely related to the number of citations as we can observe. Top-ranked institutions all correspond to well-known universities or research lab with long tradition of excellence in physics, corroborating our hypothesis that C is a reasonable proxy for ranking

Many institutions are small with few citations, hence they account for very small portion of the data. For the rest of the chapter, we will focus on the thousand most cited institutions, accounting for more than 99% of papers. They correspond to institutions with at least 698 citations within the APS data over the 120-year period (shaded area in Fig. 8.2B).

The strong correlations between the three quantities (A, P, C) indicate any of the three could characterise an institution, serving as a proxy of its ranking against others. Here, we choose C (the total number of citations) as our parameter to approximate the ranking by reputation. Other parameters such as the h-index of an institution or the number of papers P could also be used [57, 359, 360]. But the results should be insensitive to this choice owing to good correlations between these quantities ($R^2 = 0.96$ and $R^2 = 0.92$ respectively). The top-ranked institutions all correspond to well-known universities or research labs with long tradition of excellence in physics (Fig. 8.4), corroborating our hypothesis that C is a reasonable proxy for ranking. We can also observe the similarity and stability of other rankings when comparing with other metrics.

8.4 Stratification of movements and scientific impact

Thanks to the large disambiguated data spanning the last 120 years that we have compiled, a systematic study of scientific mobility is now possible. We focus on authors with similar career longevity, restricting our corpus to those who began their career between 1950 and 1980 and published for at least 20 years without any interruption exceeding 5 years. Following these criteria, we arrived at a subset of 2,725 scientists to investigate mobility patterns and their impact on individual careers. A total of 5,915 career movements are detected for this corpus.

In Figure 8.5A we select three individuals as exemplary career histories. Each line represents one individual, with circles denoting his/her publications, allowing us to observe his/her location. The size of the circle is proportional to citations the paper acquires in five years, approximating the impact of the work. By studying the whole corpus, we compute $P(m)$, the probability for a scientist to have visited m different institutions along his career (Fig. 8.5C), finding that career movements are common but infrequent: Only 14% of them never moved at all ($m = 1$). For the ones that moved, they mostly moved once or twice, $P(m)$ decaying quickly as m increases. We also compute $P(t)$, the probability to observe a movement at time t, where $t = 0$ corresponds to the date of the scientist's first publication. We find that most movements occurred in the early stage of the career (Fig. 8.5B), supporting the hypothesis that changing affiliations is a rite of passage for young researchers [347]. This likely corresponds to the postdoc period where graduates broaden their horizons through mobility. This may also reflect the increasing cost of relocation and family constraints as family developed [346, 348]. A third characteristic is the geographical distance of movements, Δd. Existing literature hints for somewhat competing hypothesis in the role geography plays in career movements. Indeed, research on human mobility suggests that regular human movements mostly cover short distances with occasional longer trips, characterized by a power law distance distribution [96, 97, 114, 282]; in contrast, country-level surveys find increasing cross-country movements mostly due to cultural exposure and life quality concerns, indicating potential dominance in long distance moves in career choices comparing with typical human travels [344–346, 348, 361–363]. We measure the distance distribution over all moves observed in our dataset, finding that our result is supported by a combination of both hypothesis.

We find the probability to move to further locations decays as a power law [364, 365] , whereas the null model predicts this probability to be flat (Fig. 8.5d). This observation is consistent with studies on human mobility, that short distance moves dominate career choices. Yet, when comparing the power law exponents, we find the exponent characterizing career moves ($\gamma = 0.65 \pm 0.053$) is much smaller than those observed in human travel ($\gamma \approx 2$), corresponding to higher likelihood of observing long range movements. This observation might be explained by the influence that scientific collaborations can have on career movements as similar low exponents are observed for collaboration network between cities [41].

Taken together, the preceding results indicate that career moves mostly happen during the early stage of a career and are more likely to cover short distances. The observed location in both time and space raises the question of how individuals move as a function of institutional rankings. To this end, denoting with $T_{i,j}$ the number of transitions from the institution of rank i to the one of rank j, we measure $P(i, j)$, the probability to have a transition from rank i to rank j as

$$P(i, j) = \frac{T_{i,j}}{\sum_{i,j} T_{i,j}}. \tag{8.1}$$

Interestingly, we find that most movements involve elite institutions (rank is small), and transitions between bottom institutions are rare (Fig. 8.6A). This is due to the fact that elite institutions are characterised by larger populations, hence translating into more events.

To account for the population based heterogeneity, we compare the observed $P(i, j)$ with the probability $P^{null}(i, j)$ expected in a random model where we randomly shuffle the transitions from institution i to j while preserving the total number of transitions from and to each institution. Formally, in this null model, we have

$$P^{null}(i, j) = \sum_k P(k, j) \cdot \sum_l P(i, l), \tag{8.2}$$

and we compare $P(i, j)$ with the null model by computing the matrix

$$M(i, j) = \frac{P(i, j)}{\sum_k P(k, j) \sum_l P(i, l)}. \tag{8.3}$$

$M(i, j)$ is the ratio between the probability $P(i, j)$ to have a transition from rank i to j divided by the probability $P^{null}(i, j)$ when the movements are shuffled, measuring the likelihood for a move to take place by

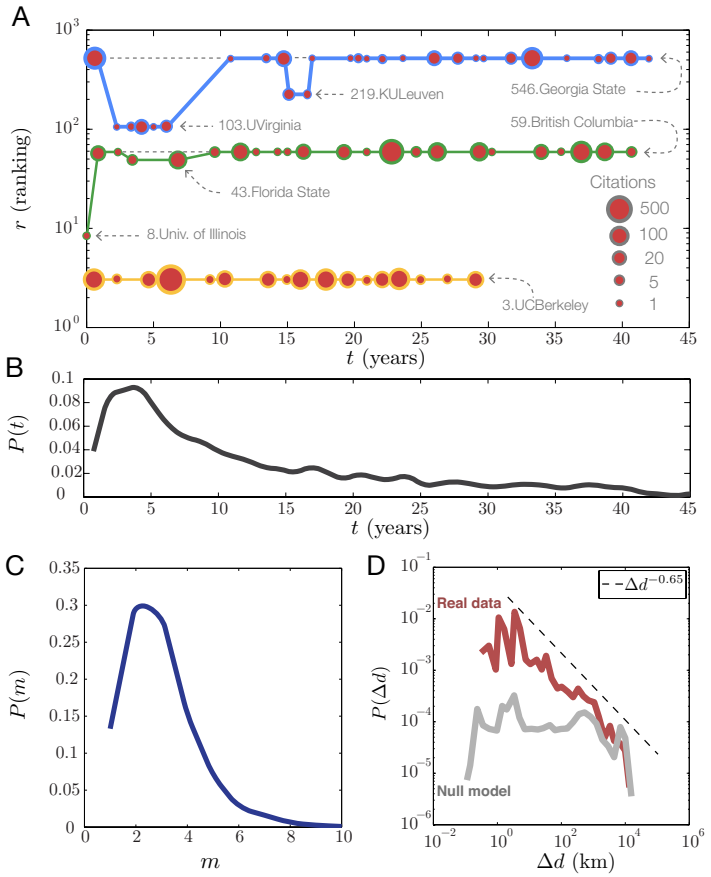


Figure 8.5: **Basic features of scientists career. (A)** Illustration of three scientific trajectories based on publications where each line corresponds to one scientist and each publication is represented by a circle whose size is proportional to its number of citations cumulated within 5 years after its publication. The institutions are ranked according to the total number of citations they obtained (see Methods), 1 being the most cited institution. **(B)** The probability density function of movement according to time, $P(t)$, shows that most movements occurred in the early stage of the career. This likely corresponds to the postdoc period where graduates broaden their horizons through mobility. **(C)** The probability density function of number of visited institutions for a scientist along his career, $P(m)$, indicates that career movements are common but infrequent. Scientists mostly move once or twice, $P(m)$ decaying quickly as m increases. **(D)** The probability density function of distance of movements, $P(\Delta d)$, has a fat-tail that can be fitted by a power law with an exponent $\gamma = 0.65 \pm 0.053$, whereas the null model predicts this probability to be roughly flat.

accounting for the size of the institutions. Hence, $M(i,j) = 1$ indicates the amount of observed movements is about what one would expect if movements were random. Similarly, $M(i,j) > 1$ indicates that we observe more transitions from i to j than we expected, whereas $M(i,j) < 1$ corresponds to transitions that are underrepresented. We find that career moves are characterized by a high degree of stratification in institutional rankings (Fig. 8.6B). Indeed, we observe two distinct clubs (red spots in Fig. 8.6B), indicating that the overrepresented movements are the ones within elite institutions (lower-left corner) or within lower-rank institutions (upper-right corner), and scientists belonging to one of the two groups tend to move to institutions within the same group. On the other hand, both upper-left and lower-right corners are colored blue, indicating cross group movements (transitions from elite to lower-rank institutions and vice-versa) are significantly underrepresented. Also, scientists from medium-ranked institutions move to the next institution with a probability that is indistinguishable from the random case. In other words, their movements indicate no bias towards middle, elite or lower-ranked institutions.

The high intensity of stratification in career movements raises an interesting question: how does individual impact in science relate to their moves across different institutional rankings ?

To answer this question, we need to quantify the impact change for each individual before and after the move. Imagine that a scientist moves from i to j, and published n papers at location i and m papers at j. The impact of a paper k can be approximated by c_k, the number of citations cumulated within 5 years after its publication [55, 143, 292, 351]. Let $c^- = \{c_1^-, c_2^-, ..., c_n^-\}$ and $c^+ = \{c_1^+, c_2^+, ..., c_m^+\}$ be the lists of number of citations for papers published before (c^-) and after (c^+) the transition from i to j ($T_{i,j}$). To quantify the change in impact, we introduce

$$\Delta c^* = \frac{\overline{c^+} - \overline{c^-}}{\sigma_c} \tag{8.4}$$

where $\overline{c^+}$ and $\overline{c^-}$ are the average of c^+ and c^-, respectively, and σ_c corresponds to the standard deviation of the concatenation of both c^+ and c^- while preserving the moment when the movement took place. Therefore, Δc^* captures the statistical difference in the average citations between papers published before and after the movement normalized by the random expectation when the same author's publications are shuffled. A positive Δc^* indicates papers following the move on average result in higher citation impact, hence representing an improvement in

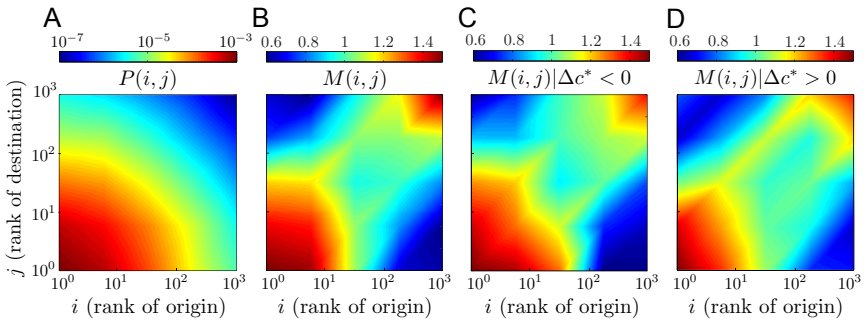


Figure 8.6: **Stratification of career movement. (A)** The matrix of probability to have a transition from rank i to rank j, (1 being the top institution) indicates that most movements involve elite institutions (rank is small) while transitions between bottom institutions are rather rare. **(B)** The likelihood $M(i,j)$ for a move to take place by accounting for the size of the institutions is characterized by a high degree of stratification in institutional rankings. Indeed, we observe two distinct clubs (red regions), indicating that the overrepresented movements are the ones within elite institutions (lower-left corner) or within lower-rank institutions (upper-right corner), and scientists belonging to one of the two groups tend to move to institutions within the same group. **(C)-(D)** The Likelihood $M(i,j)|\Delta c^* < 0$ and $M(i,j)|\Delta c^* > 0$ for transitions resulting in higher and lower scientific impact, respectively, indicates that the stratification in career moves is robust against individual performance. We find the red region in lower-left corner is more concentrated in Fig. 8.6d than in c, hinting that being more mobile in the space of rankings may lead to variable performance.

scientific impact. A negative value corresponds to a decline in impact.

To quantify the influence of movements on individual impact, we divide all movements into two categories based on the impact change: movements associated with positive and negative Δc^*. We then measure $M(i, j|\Delta c^* > 0)$ and $M(i, j|\Delta c^* < 0)$. We find the observed stratification in career moves is robust against individual impact (Fig. 8.6CD). That is, the two clubs emerge for both categories in a similar fashion as in Figure 8.6B, indicating the pattern of moving within elite or lower-rank institutions is nearly universal for people whose impact is improved or decreased following the move. Comparing Figure 8.6C and Figure 8.6D, we find the red spot in the lower-left corner is more concentrated in Figure 8.6D than in Figure 8.6C, hinting that being more mobile in the space of rankings may lead to variable impact. To test this hypothesis, for each transition $T_{i,j}$ we calculate the rank difference between the origin and destination ($\Delta r_{ij} = i - j$).

A positive value of Δr_{ij} indicates $i > j$, hence a movement to a lower-rank institution, whereas $\Delta r_{ij} < 0$ corresponds to transitions into institutions with a higher rank. In Figure 8.7 we measure the relation between Δc^* and Δr. When scientists move to institutions with a lower rank ($\Delta r > 0$), we find that their average change in impact is negative, corresponding to a decline in the impact of their work. Yet, what is particularly interesting lies in the $\Delta r < 0$ regime. Indeed, when people move from lower rank location to elite institutions, we observe no impact change on average. This is rather unexpected, as transitioning from lower-rank institutions to elite institutions is thought to provide better access to ideas and lab resources, which in turn should fuel scientific productivity. A possible explanation may be that scientist who have the opportunity to make big jumps in the ranking space may have already had an excellent impact in their previous institutions. Such a move therefore will not affect their impact.

8.5 Conclusion

In summary, we extracted affiliation information from the publications of each scientist, allowing us to reconstruct their career moves between different institutions as well as the body of work published at each location. We find career movements are common yet infrequent. Most people move only once or twice, and usually in the early stage of their

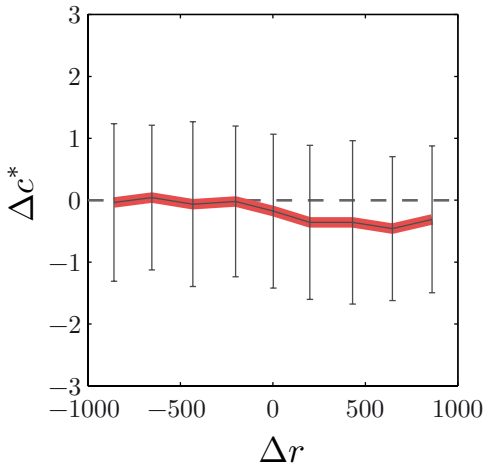


Figure 8.7: **Impact of movements on career performance.** The relation between the statistical difference of citations (Δc^*) and the ranking difference (Δr) associated to a transition shows that, when people move to institutions with a lower rank ($\Delta r > 0$), their average change in performance is negative, corresponding to a decline in the impact of their work. Yet, what is particularly interesting lies in the $\Delta r < 0$ regime. Indeed, when people move from lower rank location to elite institutions, we observe no performance change on average.

career. Career movements are affected by geography. The distance covered by the move can be approximated with a power law distribution, indicating that most movements are local and moving to faraway locations is less probable. We also observe a high degree of stratification in career movements. People from elite institutions are more likely to move to other elite institutions, whereas people from lower rank institutions are more likely to move to places with similar ranks. We further confirm that the observed stratification is robust against the change in individual performance before and after the move. When cross-group movement occurs, we find that while going from elite to lower-rank institutions on average results in a modest decrease in scientific impact, transitioning into elite institutions, does not result in gain in impact.

The nature of our dataset restricted our study on a sample of scientists. As a result of this selection process, our results are biased towards physicists from 1960s to 1980s with high career longevity. Yet, these limitations also suggest new avenues for further investigations. Indeed, as datasets become more comprehensive and of higher resolution, newly available data sources like Web of Science or Google Scholar can provide new and deeper insights towards generalization of the results across different disciplines, temporal trends, and more. Further investigations regarding the influence of career longevity on scientific mobility should also be considered as it could reveal as well results of importance. Taken together these results offer the first systematic empirical evidence on how career moves affect scientific performance and impact.

Chapter 9

General conclusions

We conclude this thesis by a short summary of the main scientific contributions presented in this work but also by highlighting the current and potential applications that can be derived from these contributions. We then discuss further research directions that are well in line with the results developed in this thesis and that are either under current investigation or worth of interest in the future. Finally, we end this thesis by a more personal discussion on the perspectives of the data revolution.

Contributions
Over the last decades, the emergence of large-scale data have unambiguously transformed many research areas from physics and biology to computer science and economics. While tremendous advances have been made in these scientific fields, much less was known about the dynamics characterising large-scale social systems until recently. Indeed, over the last few years, the increasing availability in both the scope and scale of large-scale social data capturing our everyday actions has provided fertile territory to study the underlying mechanisms behind human mobility, social interactions and individual success at an unprecedented scale. In this thesis, we provide crucial additional insights on these mechanisms, demonstrating at the same time the large interdisciplinarity associated to the study of large-scale social data and its importance to resolve crucial challenges that our society is now facing.

We start our investigation of social dynamics by analysing large-scale mobile phone data. We first investigate in Chapter 3 to what extent mobile phone calls offer reliable estimates of population distributions over entire countries, distinguishing ourselves from previous research which focussed on urban areas only. We show that not only can population maps of comparable accuracy to census data and existing downscaling methods in geography be constructed solely from mobile phone data, but that these data offer additional benefits in terms of the measurements

of population dynamics. When geographical origins and destinations of phone calls are known, spatial social networks incorporating social interactions between mobile phone users can be constructed from these data. In Chapter 4, we study the spatial structure of these social networks by computing for the first time the social communities of mobile phone users over the entire region of France. Not only we show the unexpected social influence of administrative regions in France, but we also introduce a new sensitivity measure that quantifies the spatial stability of these communities with a high resolution. In Chapter 5, we close our investigation on mobile phone data by studying the interplay between social interactions and human mobility. By extracting social and mobility fluxes from phone call data, we uncover a scaling relationship between the exponents characterising communication and mobility patterns, allowing us to derive one quantity from the other and hinting for a deeper connection among all systems where space plays a role.

In the second part of this thesis, we pursue our analysis of social dynamics by investigating the mechanisms behind success based on publication data. We first present in Chapter 6 three techniques that aim at disambiguating three common types of ambiguous information present in these digital libraries; Author names, affiliations and topics. Using these techniques to obtain a disambiguated set of scientific careers, we study in Chapter 7 the patterns of productivity and impact associated to individual scientists. While we demonstrate the existence of reproducible productivity patterns leading to the highest impact work of a scientist, we also show that highly cited articles appear to be dramatically unpredictable in a scientist's career. We also highlight peculiar statistical features associated to scientists with outstanding publications. We close our investigation of success dynamics by quantifying in Chapter 8 the effect of mobility on scientific impact. By tracking affiliation information from publications we first show how to reconstruct career trajectories of individual scientists over decades. We then show that scientific mobility is characterised by a high degree of stratification; People from elite institutions are more likely to move to other elite institutions, whereas people from lower rank institutions are more likely to move to places with similar ranks. We further confirm that the observed stratification is robust against the change in individual impact before and after the move. Finally, when cross-group movement occurs, we demonstrate that while going from elite to lower-rank institutions on average results in a modest decrease in scientific impact, transitioning into elite institutions, does not result in gain in impact. Taken together these results offer the first systematic empirical evidence on how career moves affect scientific

productivity and impact.

Applications
While the different results presented in this work bring novel and valuable understandings about the dynamics characterising social systems, we also demonstrate in this thesis that concrete applications can be derived from these insights.

In chapter 3, our ability to translate mobile phone activities into detailed population densities enables us to design a cost-efficient method to map population over large geographical extent over time. While we highlight its efficiency to detect dynamical changes in developed countries, its principal application remains for low-income countries which are the most vulnerable to disasters, outbreaks or conflicts and where few or no information regarding population distribution over time is available. The ability of our method to obtain spatially and temporally detailed population distributions can potentially provide the essential denominator required in many fields, such as studying collective human responses to disease outbreaks or emergencies, assessing vulnerabilities, calculating populations at risk of human or natural disasters or deriving health and development indicators.

In Chapter 4, the observed correspondance between social communities and administrative regions, which have until now been viewed as no more than a collection of culturally distinct departments, would seem to indicate that these administrative limits confer a deeply rooted sense of shared identity to the population within each region. This observation demonstrates the importance of integrating large-scale social data in political processes in order to minimise social, political or ethnic conflict in particular situations. A good example remains the recent redraw of administrative regions of France, which led to long-standing social and political conflicts in 2014. Another example includes the anecdotical evidence that suggests that some departments were originally associated to one region over another based on the frequency of landline communications between population [366]. Furthermore, the sensitivity measure introduced in that same chapter can also reveal additional important details about the social and spatial stability of particular elements present in a social network. This measure can not only be useful in political processes but also in business applications such as churn prediction where assessing the sensitivity of particular costumers towards particular products remains a challenge.

As illustrated in Chapter 5, the identified scaling law between social interactions and human mobility also offers concrete applications for social systems. Indeed, our ability to derive mobility information from social fluxes, i.e. phone calls that can be readily extracted by network providers, opens up a new avenue not only for traffic forecasting applications but also for urban planning applications and epidemic spreading as demonstrated in this work. While we focussed on mobile phone data, there exist other valuable sources of social interactions in order to derive mobility patterns at a very large scale. Indeed, social network platforms, such as Facebook or Twitter for example, often contain geotagged comments or tweets, enabling the possibility to derive social but also mobility fluxes of individuals. Understanding and modelling that same relationship between social and mobility fluxes from social media data not only represents an exciting research question worth investigating, but it could also lead to promising mobility prediction applications not only valuable for traffic prediction or epidemic spreading, but also for marketeers when coupled with social media marketing campaigns.

Besides phone call data, the techniques introduced in Chapter 6 demonstrate that other types of large-scale social data, namely publication data, can also be mined to derive practical applications for science. First, disambiguating author names can fuel applications where individual careers need to be reconstructed. This technique, coupled with our analysis of individual productivity and impact developed in Chapter 7, could offer for example an interesting tool for hiring procedures in science where current individual productivity and impact measures are often misleading. Similarly, disambiguating affiliations also provides valuable insights about unknown mobility patterns in science, as illustrated in Chapter 8, which remain crucial for policy-makers. Finally, the automated topic detection method could also be integrated to online search and repository tools such as Google Scholar or Arxiv platforms to improve not only automated document classification but also personal user recommendation.

Beyond these applications dedicated to science, the techniques developed in Chapter 6 and illustrated in Chapters 7 and 8 offer promising solutions for concrete business applications. Indeed, as we are now overwhelmed by social data originating from myriads of sources, the need to reconnect online profiles to customers or particular topics to products has become a serious challenge for companies. These techniques, if translated to user/customer networks, can address these issues and can provide crucial solutions to customer insights problematics.

Further Work
While we presented scientific contributions corresponding to well defined or completed research projects, several questions well in line with this thesis are still under investigation. We provide here a description of these different research questions.

As we mentioned earlier, the main objective of the population mapping method developed in Chapter 3 is to help organisations tackle human disasters in low-income countries by providing accurate population distributions over time. To that end, two distinct research projects incorporating this novel method have been initiated. The first one, conducted in collaboration with the WorldPop project, consists of collecting call data records from Namibia to further test the extrapolation capacity and sensitivity of our method in data scarce regions. For the second project, we are investigating the possibility to incorporate drug purchase data as additional input to the method in order to provide accurate predictions of viral disease spreading in a given region over time. This second project results from a collaboration with a major telecommunication operator and a renowned health institute.

Motivated by the accuracy of the disambiguation techniques introduced in Chapter 6, an additional project investigating the mechanisms behind success has also been initiated. As individual careers spanning several decades as well as research topics can now be derived from digital librairies, one can analyse the different *footprints* left by particular scientists in terms of productivity, collaboration patterns and topic changes over time. This ability to gather detailed information about individual careers over time raises several important questions: are there particular categories of scientists in terms of topic changes, collaboration patterns or productivity ? If there are, how do these categories relate to individual success ? Are there better scientific strategies than others ? This ongoing project aims at addressing these questions.

In science, anecdotal evidence exists that high-achievers are often protégés of illustrious mentors [367–369]. While projects like the mathematical genealogy document clear signs of such chaperone bonds between renowned scientists [54], we lack systematic quantitative evidence of the role of apprenticeship in scientific publishing and, in general, of how scientific knowledge is passed down between different generations of scientists [54, 370]. Our last ongoing project aims at quantifying this chaperone phenomenon in science by leveraging the detailed information

that are now available through large-scale publication data. While this project is still under investigation, we find however that the chaperone phenomenon is well established in science and has become more pronounced in the last decade, with different magnitude depending on the research field.

Where is this going ?

"*You can't manage what you can't measure*" as W. Edwards Deming used to say. Today, I couldn't agree more as this really illustrates the importance of the recent data explosion. This *data revolution* is impacting not only marketeers but all major actors in all industries. In health care, data sources such as medical and insurance records, wearable sensors, genetic data and social media use are already used to draw a comprehensive picture of the patient. As prevention is better than cure, we can imagine the tremendous social but mainly commercial opportunities associated to such transformations; tailored healthcare package, drug test efficiency, smarter screening, and even claims fraud detection. In Astronomy, data from hundreds of satellites as well as from thousands of radio telescopes are also aggregated and analysed. Such rich amount of data can be used for very different purposes, from locating a single rare star in a haystack of billions of near-identical stars to understanding the spatial correlation of every single galaxy in the universe. In the music industry, data gathered from downloads, sales, social comments enable marketeers to design *data-driven ears* that can understand, analyse and predict success of particular songs at an early stage. Finally, in the automotive industry, tremendous volume and variety of data are generated but also processed and analysed. Such data can either originate from the car itself through sensors and cameras or either originate from external sources such as traffic and weather alerts to offer a better driving experience but also improve safety.

As we can see, this *Big Data* revolution rapidly growing in popularity offers promises that would drastically change most aspects of our lives. While this thesis aims at demonstrating in some aspects what we *can* do and achieve with these data, I believe this work bears also a responsibility to discuss what we *can't* do as well as the pitfalls and dangers associated to this data explosion, precisely because of this newfound popularity and growing use.

Throughout my thesis, I have too often experienced being overwhelmed by the increasing amount of data available. While data is important, the right data is essential. As datasets are growing more

and more complex, it is thus crucial to dedicate a significant amount of time to the processing and understanding of these data in order to extract the right dataset for your project or product. While volume cannot be neglected, data reliability remains indeed the most important aspect associated to it and quantity of data does not mean that one can ignore fundamental issues related to measurements. One popular failure illustrating this issue remains the (un)popular Google Flu Trends (GFT) project. Introduced in 2009, this project aimed at predicting and detecting the spread of flu as accurately and quickly as authorities, based on flu-related search queries. Surprisingly, a few years later, GFT appeared to make more bad predictions than correct ones [371]. Scientists attributed this failure to the reliability of the data themselves [372]; collections of data relying on web hits are often risky to merged as they might be collected in different ways. The patterns in data collected at a particular time do not necessarily apply to data collected at another time. Reliability and robustness of data are thus a key success factor in research but also in any other industry. Moreover, data should not be seen as a substitute to traditional methods, it is an adjunct to scientific inquiry, a complement, that too many actors in science and in the industry often forget.

If data reliability and robustness of analysis remain a crucial challenge both for data science and for many industries, another battle has already been fought and lost: the battle for privacy. Today, personal data are routinely collected and traded in the new economy. We have become the product. The popularity around big data is attracting millions of dollars from investors and brands, hoping to turn a profit out of personal data. At the same time, governments through intelligent agencies are also massively collecting personal activities but for much different purposes. Moreover, emerging technologies like connected cars and glasses or facial recognition are perfect ingredients for ubiquitous large-scale online surveillance.

At the same time, this big data revolution is also a source for discrimination. Today, big data analytics affect many things from employment and promotions to fair housing and loans. Predictive analysis by the public or private sector can now be used to make determinations about our ability to get a job, a loan, a credit card or even to fly. We fought for decades against discrimination, but big data analytics enable one to make it essentially automated.

Today, privacy researcher and analysts are calling for legislations

to regulate the collection, storage and trade of personal data. This is of course far from being trivial as many companies operate on different continents and countries, each having its own jurisdiction and regulation. We would thus need enough people across very different places to urge their governments and companies to change their privacy position. But our worst enemy in this battle might be simply ourselves; How can we expect people to fight something they don't know or understand? For example, how many times did you read the terms of service or privacy policy before installing a mobile phone application or subscribing to a service? Our first act should thus be to raise awareness and educate ourself and others about privacy. I truly believe this *Big Data* revolution is important and positive for our lives and it should be. But we need to stay aware of its dangers and potential deflections when it comes to personal data so that we can better initiate and support regulations around this revolution.

Bibliography

[1] Seref Sagiroglu and Duygu Sinanc, "Big data: A review", in: *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, IEEE, 2013, pp. 42–47.

[2] John Gantz and David Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", in: *IDC iView: IDC Analyze the Future* 2007 (2012), pp. 1–16.

[3] Adam Jacobs, "The pathologies of big data", in: *Communications of the ACM* 52.*8* (2009), pp. 36–44.

[4] Martin Hilbert and Priscila López, "The world?s technological capacity to store, communicate, and compute information", in: *science* 332.*6025* (2011), pp. 60–65.

[5] Hsinchun Chen, Roger HL Chiang, and Veda C Storey, "Business Intelligence and Analytics: From Big Data to Big Impact.", in: *MIS quarterly* 36.*4* (2012), pp. 1165–1188.

[6] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz, "Big data, analytics and the path from insights to value", in: *MIT Sloan Management Review* 21 (2013).

[7] Pål Sundsøy, Johannes Bjelland, Asif M Iqbal, Yves-Alexandre de Montjoye, et al., "Big Data-Driven Marketing: How machine learning outperforms marketers? gut-feeling", in: *Social Computing, Behavioral-Cultural Modeling and Prediction*, Springer, 2014, pp. 367–374.

[8] Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al., "Big data: The future of biocuration", in: *Nature* 455.*7209* (2008), pp. 47–50.

[9] Elena Aronova, Karen S Baker, and Naomi Oreskes, "Big science and big data in biology: from the international geophysical year through the international biological program to the long term ecological research (LTER) network, 1957–present", in: (2010).

[10] Vivien Marx, "Biology: The big challenges of big data", in: *Nature* 498.*7453* (2013), pp. 255–260.

[11] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters", in: *Communications of the ACM* 51.*1* (2008), pp. 107–113.

[12] Liran Einav and Jonathan D Levin, *The data revolution and economic analysis*, tech. rep., National Bureau of Economic Research, 2013.

[13] Hal R Varian, "Big data: New tricks for econometrics", in: *The Journal of Economic Perspectives* (2014), pp. 3–27.

[14] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al., "Life in the network: the coming age of computational social science", in: *Science (New York, NY)* 323.*5915* (2009), p. 721.

[15] Claudio Cioffi-Revilla, "Computational social science", in: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.*3* (2010), pp. 259–271.

[16] Carlo Ratti, S Williams, D Frenchman, and RM Pulselli, "Mobile landscapes: using location data from cell phones for urban analysis", in: *Environment and Planning B Planning and Design* 33.*5* (2006), p. 727.

[17] R Pulselli, P Ramono, Carlo Ratti, and E Tiezzi, "Computing urban mobile landscapes through monitoring population density based on cellphone chatting", in: *Int. J. of Design and Nature and Ecodynamics* 3.*2* (2008), pp. 121–134.

[18] Jonathan Reades, Francesco Calabrese, and Carlo Ratti, "Eigenplaces: analysing cities using the space-time structure of the mobile phone network", in: *Environment and Planning B: Planning and Design* 36.*5* (2009), pp. 824–836.

[19] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava, "Using mobile phones to determine transportation modes", in: *ACM Transactions on Sensor Networks (TOSN)* 6.*2* (2010), p. 13.

[20] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson, "VTrack: accurate, energy-aware road traffic delay estimation using mobile phones", in: *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, ACM, 2009, pp. 85–98.

[21] Hui Zheng, Dong Li, and Zhuo Gao, "An epidemic model of mobile phone virus", in: *Pervasive Computing and Applications, 2006 1st International Symposium on*, IEEE, 2006, pp. 1–5.

[22] Pu Wang, Marta C González, César A Hidalgo, and Albert-László Barabási, "Understanding the spreading patterns of mobile phone viruses", in: *Science* 324.*5930* (2009), pp. 1071–1076.

[23] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza, "On the use of human mobility proxies for modeling epidemics", in: *PLoS computational biology* 10.*7* (2014), e1003716.

[24] G. Krings, F. Calabrese, C. Ratti, and V.D. Blondel, "Urban gravity: a model for inter-city telecommunication flows", in: *Journal of Statistical Mechanics: Theory and Experiment* 2009 (2009), p. L07003.

[25] Markus Schläpfer, Luís MA Bettencourt, Sébastian Grauwin, Mathias Raschke, Rob Claxton, Zbigniew Smoreda, Geoffrey B West, and Carlo Ratti, "The scaling of human interactions with city size", in: *Journal of The Royal Society Interface* 11.*98* (2014), p. 20130789.

[26] Vincent Blondel, Gautier Krings, Isabelle Thomas, et al., "Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone", in: *Brussels Studies* 42.*4* (2010), pp. 1–12.

[27] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz, "Redrawing the map of Great Britain from a network of human interactions", in: *PloS one* 5.*12* (2010), e14248.

[28] Vincent Blondel, **Pierre Deville**, Frédéric Morlot, Zbigniew Smoreda, Paul Van Dooren, Cezary Ziemlicki, et al., "Voice on the Border: Do Cellphones Redraw the Maps?", in: *ParisTech Review* (2011).

[29] Andrew Tatem and Catherine Linard, "Population mapping of poor countries", in: *Nature* 474.*7349* (2011), pp. 36–36.

[30] John Bongaarts and Steven Sinding, "Population policy in transition in the developing world", in: *Science* 333.*6042* (2011), pp. 574–576.

[31] Andrew J Tatem, Andres J Garcia, Robert W Snow, Abdisalan M Noor, Andrea E Gaughan, Marius Gilbert, and Catherine Linard, "Millennium development health metrics: where do Africa's children and women of childbearing age live?", in: *Population health metrics* 11.*1* (2013), p. 11.

[32] Francesco Checchi, Barclay T Stewart, Jennifer J Palmer, and Chris Grundy, "Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations", in: *Int J Health Geogr* 12.*4* (2013).

[33] Catherine Linard and Andrew J Tatem, "Large-scale spatial population databases in infectious disease research", in: *Int J Health Geogr* 11.*7* (2012).

[34] Brian C O'Neill, Michael Dalton, Regina Fuchs, Leiwen Jiang, Shonali Pachauri, and Katarina Zigova, "Global demographic trends and future carbon emissions", in: *Proceedings of the National Academy of Sciences* 107.*41* (2010), pp. 17521–17526.

[35] John O?Loughlin, Frank DW Witmer, Andrew M Linke, Arlene Laing, Andrew Gettelman, and Jimy Dudhia, "Climate variability and conflict risk in East Africa, 1990–2009", in: *Proceedings of the National Academy of Sciences* 109.*45* (2012), pp. 18344–18349.

[36] August Lösch, "The nature of economic regions", in: *Southern Economic Journal* (1938), pp. 71–78.

[37] Masahisa Fujita, Paul R Krugman, and Anthony J Venables, *The spatial economy: Cities, regions, and international trade*, MIT press, 2001.

[38] Michael Storper and Anthony J Venables, "Buzz: face-to-face contact and the urban economy", in: *Journal of economic geography* 4.*4* (2004), pp. 351–370.

[39] Mark EJ Newman, "The structure of scientific collaboration networks", in: *Proceedings of the National Academy of Sciences* 98.*2* (2001), pp. 404–409.

[40] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek, "Evolution of the social network of scientific collaborations", in: *Physica A: Statistical mechanics and its applications* 311.*3* (2002), pp. 590–614.

[41] Raj Kumar Pan, Kimmo Kaski, and Santo Fortunato, "World citation and collaboration networks: uncovering the role of geography in science", in: *Scientific reports* 2 (2012).

[42] Floriana Gargiulo and Timoteo Carletti, "Driving forces of researchers mobility", in: *Scientific reports* 4 (2014).

[43] Hua-Wei Shen and Albert-László Barabási, "Collective credit allocation in science", in: *Proceedings of the National Academy of Sciences* 111.*34* (2014), pp. 12325–12330.

[44] Jonathan Stallings, Eric Vance, Jiansheng Yang, Michael W Vannier, Jimin Liang, Liaojun Pang, Liang Dai, Ivan Ye, and Ge Wang, "Determining scientific impact using a collaboration index", in: *Proceedings of the National Academy of Sciences* 110.*24* (2013), pp. 9680–9685.

[45] Nils T Hagen, "Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis", in: *PLoS One* 3.*12* (2008), e4021.

[46] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, "Diffusion of scientific credits and the ranking of scientists", in: *Physical Review E* 80.*5* (2009), p. 056103.

[47] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance.", in: *Psychological Review* 100.*3* (1993), p. 363.

[48] K Anders Ericsson, Michael J Prietula, and Edward T Cokely, "The making of an expert", in: *Harvard Business Review* 85.*7/8* (2007), p. 114.

[49] Alexander M Petersen, Massimo Riccaboni, H Eugene Stanley, and Fabio Pammolli, "Persistence and uncertainty in the academic career", in: *Proceedings of the National Academy of Sciences* 109.*14* (2012), pp. 5213–5218.

[50] Alexander M Petersen, Santo Fortunato, Raj K Pan, Kimmo Kaski, Orion Penner, Massimo Riccaboni, H Eugene Stanley, and Fabio Pammolli, "Reputation and impact in academic careers", in: *arXiv preprint arXiv:1303.7274* (2013).

[51] Nathan Eagle and Alex Pentland, "Reality mining: sensing complex social systems", in: *Personal and ubiquitous computing* 10.*4* (2006), pp. 255–268.

[52] Emiliano Miluzzo, Nicholas D Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B Eisenman, Xiao Zheng, and Andrew T Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application", in: *Proceedings of the 6th ACM conference on Embedded network sensor systems*, ACM, 2008, pp. 337–350.

[53] Arnout van de Rijt, Soong Moon Kang, Michael Restivo, and Akshay Patil, "Field experiments of success-breeds-success dynamics", in: *Proceedings of the National Academy of Sciences* 111.*19* (2014), pp. 6934–6939.

[54] R Dean Malmgren, Julio M Ottino, and Luís A Nunes Amaral, "The role of mentorship in protégé performance", in: *Nature* 465.*7298* (2010), pp. 622–626.

[55] Dashun Wang, Chaoming Song, and Albert-László Barabási, "Quantifying long-term scientific impact", in: *Science* 342.*6154* (2013), pp. 127–132.

[56] David van Dijk, Ohad Manor, and Lucas B Carey, "Publication metrics and success on the academic job market", in: *Current Biology* 24.*11* (2014), R516–R517.

[57] Jorge E Hirsch, "An index to quantify an individual's scientific research output", in: *Proceedings of the National academy of Sciences of the United States of America* 102.*46* (2005), pp. 16569–16572.

[58] Matthew J Salganik and Duncan J Watts, "Social influence: the puzzling nature of success in cultural markets", in: *P. Hedström and P. Bearman* (2009), pp. 315–341.

[59] **Pierre Deville**, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem, "Dynamic population mapping using mobile phone data", in: *Proceedings of the National Academy of Sciences* 111.*45* (2014), pp. 15888–15893.

[60] **Pierre Deville**, Vincent D Blondel, Paul Van Dooren, and Zbigniew Smoreda, "Mobile phone communications help identify stable regions in France", in: *Proceedings of the 31th Benelux Meeting on Systems and Control*.

[61] **Pierre Deville**, Dashun Wang, Chaoming Song, Nathan Eagle, Vincent D Blondel, and Albert-László Barabási, "Scaling Identity in Spatial Networks: Connections between Mobility and Social Interactions", in: (Under review in PNAS).

[62] Roberta Sinatra, **Pierre Deville**, Dashun Wang, Michael Szell, and Albert-László Barabási, "A Century of Physics", in: (Nature Physics, October 2015).

[63] Roberta Sinatra, Dashun Wang, **Pierre Deville**, Chaoming Song, and Albert-Laszlo Barabasi, "Scientific impact: the story of your big hit", in: (Under review in Science).

[64] **Pierre Deville**, Dashun Wang, Roberta Sinatra, Chaoming Song, Vincent D Blondel, and Albert-László Barabási, "Career on the move: geography, stratification, and scientific impact", in: *Scientific reports* 4 (2014).

[65] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clerot, **Pierre Deville**, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki, "Data for Development: the D4D Challenge on Mobile Phone Data", in: *arXiv preprint arXiv:1210.0137* (2012).

[66] Vedran Sekara, Roberta Sinatra, **Pierre Deville**, Sebastian Ahnert, Albert-László Barabási, and Sune Lehmann, "The chaperone phenomenon in science", in: (in preparation).

[67] Alexander George and Andrew Bennett, *Case studies and theory development*, Free Press, 1979.

[68] Christopher H Achen and Duncan Snidal, "Rational deterrence theory and comparative case studies", in: *World Politics* 41.*02* (1989), pp. 143–169.

[69] Stanley Lieberson, "Small N's and big conclusions: an examination of the reasoning in comparative studies based on a small number of cases", in: *Social forces* 70.*2* (1991), pp. 307–320.

[70] Gary King, Robert O Keohane, and Sidney Verba, *Designing social inquiry: Scientific inference in qualitative research*, Princeton University Press, 1994.

[71] Vincent D Blondel, Adeline Decuyper, and Gautier Krings, "A survey of results on mobile phone datasets analysis", in: *arXiv preprint arXiv:1502.03406* (2015).

[72] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi, "Understanding individual human mobility patterns", in: *Nature* 453.*7196* (2008), pp. 779–782.

[73] Jon Froehlich, Leah Findlater, and James Landay, "The design of eco-feedback technology", in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 1999–2008.

[74] Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke, Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, and Péter Boda, "PEIR, the personal environmental impact report, as a platform for participatory sensing systems research", in: *Proceedings of the 7th international conference on Mobile systems, applications, and services*, ACM, 2009, pp. 55–68.

[75] Predrag Klasnja and Wanda Pratt, "Healthcare in the pocket: mapping the space of mobile-phone health interventions", in: *Journal of biomedical informatics* 45.*1* (2012), pp. 184–198.

[76] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al., "Activity sensing in the wild: a field trial of ubifit garden", in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2008, pp. 1797–1806.

[77] Konrad Lorincz, Bor-rong Chen, Geoffrey Werner Challen, Atanu Roy Chowdhury, Shyamal Patel, Paolo Bonato, Matt Welsh, et al., "Mercury: a wearable sensor network platform for high-fidelity motion analysis.", in: *SenSys*, vol. 9, 2009, pp. 183–196.

[78] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng, "Why we twitter: understanding microblogging usage and communities", in: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, 2007, pp. 56–65.

[79] Amanda L Traud, Peter J Mucha, and Mason A Porter, "Social structure of Facebook networks", in: *Physica A: Statistical Mechanics and its Applications* 391.*16* (2012), pp. 4165–4180.

[80] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn, "Virality prediction and community structure in social networks", in: *Scientific reports* 3 (2013).

[81] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec, "Can cascades be predicted?", in: *Proceedings of the 23rd international conference on World wide web*, International World Wide Web Conferences Steering Committee, 2014, pp. 925–936.

[82] Kate Starbird and Leysia Palen, "(How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising", in: *Proceedings of the acm 2012 conference on computer supported cooperative work*, ACM, 2012, pp. 7–16.

[83] W Lance Bennett, Alexandra Segerberg, and Shawn Walker, "Organization in the crowd: peer production in large-scale networked protests", in: *Information, Communication & Society* 17.*2* (2014), pp. 232–260.

[84] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al., "The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions", in: *International journal of communication* 5 (2011), p. 31.

[85] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness", in: *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2010, pp. 1079–1088.

[86] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo, "Twitter Under Crisis: Can we trust what we RT?", in: *Proceedings of the first workshop on social media analytics*, ACM, 2010, pp. 71–79.

[87] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", in: *Proceedings of the 19th international conference on World wide web*, ACM, 2010, pp. 851–860.

[88] Mark EJ Newman, "Scientific collaboration networks. I. Network construction and fundamental results", in: *Physical review E* 64.*1* (2001), p. 016131.

[89] Aaron Clauset, Samuel Arbesman, and Daniel B Larremore, "Systematic inequality and hierarchy in faculty hiring networks", in: *Science Advances* 1.*1* (2015), e1400005.

[90] Mark W Horner and Morton E O'Kelly, "Embedding economies of scale concepts for hub network design", in: *Journal of Transport Geography* 9.*4* (2001), pp. 255–265.

[91] Ryuichi Kitamura, Cynthia Chen, Ram M Pendyala, and Ravi Narayanan, "Micro-simulation of daily activity-travel patterns for travel demand forecasting", in: *Transportation* 27.*1* (2000), pp. 25–51.

[92] Timothy J Hatton, Jeffrey G Williamson, et al., *Global migration and the world economy: Two centuries of policy and performance*, Cambridge Univ Press, 2005.

[93] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions", in: *PLoS medicine* 4.*1* (2007), e13.

[94] Lars Hufnagel, Dirk Brockmann, and Theo Geisel, "Forecast and control of epidemics in a globalized world", in: *Proceedings of the National Academy of Sciences of the United States of America* 101.*42* (2004), pp. 15124–15129.

[95] Jon Kleinberg, "Computing: The wireless epidemic", in: *Nature* 449.*7160* (2007), pp. 287–288.

[96] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel", in: *Nature* 439.*7075* (2006), pp. 462–465.

[97] M.C. González, C.A. Hidalgo, and A.L. Barabási, "Understanding individual human mobility patterns", in: *Nature* 453.*7196* (2008), pp. 779–782.

[98] C. Song, T. Koren, P. Wang, and A.L. Barabási, "Modelling the scaling properties of human mobility", in: *Nature Physics* (2010).

[99] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási, "Limits of predictability in human mobility", in: *Science* 327.*5968* (2010), pp. 1018–1021.

[100] Xin Lu, Linus Bengtsson, and Petter Holme, "Predictability of population displacement after the 2010 Haiti earthquake", in: *Proceedings of the National Academy of Sciences* 109.*29* (2012), pp. 11576–11581.

[101] George Kingsley Zipf, "The P1 P2/D hypothesis: On the intercity movement of persons", in: *American sociological review* (1946), pp. 677–686.

[102] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases", in: *Proceedings of the National Academy of Sciences* 106.*51* (2009), pp. 21484–21489.

[103] Woo-Sung Jung, Fengzhong Wang, and H Eugene Stanley, "Gravity model in the Korean highway", in: *EPL (Europhysics Letters)* 81.*4* (2008), p. 48005.

[104] Andrea De Montis, Marc Barthélemy, Alessandro Chessa, and Alessandro Vespignani, "The structure of inter-urban traffic: A weighted network analysis", in: *arXiv preprint physics/0507106* (2005).

[105] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann, "The structure of borders in a small world", in: *PloS one* 5.*11* (2010), e15422.

[106] Pablo Kaluza, Andrea Kölzsch, Michael T Gastner, and Bernd Blasius, "The complex network of global cargo ship movements", in: *Journal of the Royal Society Interface* (2010), rsif20090495.

[107] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte, "Uncovering space-independent communities in spatial networks", in: *Proceedings of the National Academy of Sciences* 108.*19* (2011), pp. 7663–7668.

[108] Samuel A Stouffer, "Intervening opportunities: a theory relating mobility and distance", in: *American sociological review* 5.*6* (1940), pp. 845–867.

[109] Kingsley E Haynes, Dudley L Poston, and Paul Schnirring, "Intermetropolitan migration in high and low opportunity areas: indirect tests of the distance and intervening opportunities hypotheses", in: *Economic Geography* (1973), pp. 68–73.

[110] Charles CHEUNG and John BLACK, "Residential location-specific travel preferences in an intervening opportunities model: transport assessment for urban release areas", in: *Journal of the Eastern Asia Society for Transportation Studies* 6 (2005), pp. 3773–3788.

[111] Jean-Michel Guldmann, "Competing destinations and intervening opportunities interaction models of inter-city telecommunication flows*", in: *Papers in Regional Science* 78.*2* (1999), pp. 179–194.

[112] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo, "A tale of many cities: universal patterns in human urban mobility", in: *PloS one* 7.*5* (2012), e37027.

[113] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky, "A tale of two cities", in: *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, ACM, 2010, pp. 19–24.

[114] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási, "A universal model for mobility and migration patterns", in: *Nature* 484.*7392* (2012), pp. 96–100.

[115] James D Murray, *Mathematical Biology I: An Introduction, vol. 17 of Interdisciplinary Applied Mathematics*, 2002.

[116] Vittoria Colizza, Marc Barthélemy, Alain Barrat, and Alessandro Vespignani, "Epidemic modeling in complex realities", in: *Comptes rendus biologies* 330.*4* (2007), pp. 364–374.

[117] Joseph SM Peiris, Kwok Y Yuen, Albert DME Osterhaus, and Klaus Stöhr, "The severe acute respiratory syndrome", in: *New England Journal of Medicine* 349.*25* (2003), pp. 2431–2441.

[118] Marcelo FC Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis Chao, Ira Longini, M Elizabeth Halloran, and Alessandro Vespignani, "Assessing the international spreading risk associated with the 2014 West African Ebola outbreak", in: *PLOS Currents Outbreaks* 1 (2014).

[119] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani, "The role of the airline transportation network in the prediction and predictability of global epidemics", in: *Proceedings of the National Academy of Sciences of the United States of America* 103.*7* (2006), pp. 2015–2020.

[120] Cécile Viboud, Ottar N Bjørnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell, "Synchrony, waves, and spatial hierarchies in the spread of influenza", in: *science* 312.*5772* (2006), pp. 447–451.

[121] Xin-Jian Xu, Xun Zhang, and JFF Mendes, "Impacts of preference and geography on epidemic spreading", in: *Physical Review E* 76.*5* (2007), p. 056109.

[122] Vitaly Belik, Theo Geisel, and Dirk Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases", in: *Physical Review X* 1.*1* (2011), p. 011001.

[123] Stefano Merler and Marco Ajelli, "The role of population heterogeneity and human mobility in the spread of pandemic influenza", in: *Proceedings of the Royal Society of London B: Biological Sciences* 277.*1681* (2010), pp. 557–565.

[124] Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Michele Tizzoni, Vittoria Colizza, and Alessandro Vespignani, "Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic", in: *PloS one* 6.*1* (2011), e16591.

[125] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee, "Quantifying the impact of human mobility on malaria", in: *Science* 338.*6104* (2012), pp. 267–270.

[126] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthélemy, "Structure of urban movements: polycentric activity and entangled hierarchical flows", in: *PloS one* 6.*1* (2011), e15923.

[127] Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti, "The geography of taste: analyzing cell-phone mobility and social events", in: *Pervasive computing*, Springer, 2010, pp. 22–37.

[128] Fabio Manfredini, Paolo Tagliolato, and Carmelo Di Rosa, "Monitoring temporary populations through cellular core network data", in: *Computational Science and Its Applications-ICCSA 2011*, Springer, 2011, pp. 151–161.

[129] Rein Ahas, Anto Aasa, Ülar Mark, Taavi Pae, and Ain Kull, "Seasonal tourism spaces in Estonia: Case study with mobile positioning data", in: *Tourism management* 28.*3* (2007), pp. 898–910.

[130] Rein Ahas, Anto Aasa, Antti Roose, Ülar Mark, and Siiri Silm, "Evaluating passive mobile positioning data for tourism surveys: An Estonian case study", in: *Tourism Management* 29.*3* (2008), pp. 469–486.

[131] Andres Kuusik, Rein Ahas, and Margus Tiru, "Analysing repeat visitation on country level with passive mobile positioning method: an Estonian case study", in: *Discussions on Estonian economic policy: Theory and practice of economic policy.* 17 (2009).

[132] Fabien Girardin, Andrea Vaccari, Alexander Gerber, Assaf Biderman, and Carlo Ratti, "Quantifying urban attractiveness from the distribution and density of digital footprints", in: (2009).

[133] Mor Naaman, "Geographic information from georeferenced social media data", in: *SIGSPATIAL Special* 3.*2* (2011), pp. 54–61.

[134] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca, "Network analysis in the social sciences", in: *science* 323.*5916* (2009), pp. 892–895.

[135] George Kingsley Zipf, "Human behavior and the principle of least effort.", in: (1949).

[136] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins, "Geographic routing in social networks", in: *Proceedings of the National Academy of Sciences of the United States of America* 102.*33* (2005), pp. 11623–11628.

[137] Jacob Goldenberg and Moshe Levy, "Distance is not dead: Social interaction and geographical distance in the internet era", in: *arXiv preprint arXiv:0906.3202* (2009).

[138] Lars Backstrom, Eric Sun, and Cameron Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity", in: *Proceedings of the 19th international conference on World wide web*, ACM, 2010, pp. 61–70.

[139] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg, "Inferring social ties from geographic coincidences", in: *Proceedings of the National Academy of Sciences* 107.*52* (2010), pp. 22436–22441.

[140] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi, "Human mobility, social ties, and link prediction", in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2011, pp. 1100–1108.

[141] Sherwin Rosen, "The economics of superstars", in: *The American economic review* (1981), pp. 845–858.

[142] Robert H Frank and Philip J Cook, *The winner-take-all society: Why the few at the top get so much more than the rest of us*, Random House, 2010.

[143] Albert-László Barabási, Chaoming Song, and Dashun Wang, "Publishing: Handful of papers dominates citation", in: *Nature* 491.*7422* (2012), pp. 40–40.

[144] Peter Hedström and Richard Swedberg, *Social mechanisms: An analytical approach to social theory*, Cambridge University Press, 1998.

[145] Abhijit V Banerjee, "A simple model of herd behavior", in: *The Quarterly Journal of Economics* (1992), pp. 797–817.

[146] W Brian Arthur, *Increasing returns and path dependence in the economy*, University of Michigan Press, 1994.

[147] R Merdoc, "The Matthew effect in science: the reward and communication systems of science are considered", in: *Science* 159 (1968), pp. 56–63.

[148] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks", in: *science* 286.*5439* (1999), pp. 509–512.

[149] Mark EJ Newman, "Clustering and preferential attachment in growing networks", in: *Physical Review E* 64.*2* (2001), p. 025102.

[150] Hawoong Jeong, Zoltan Néda, and Albert-László Barabási, "Measuring preferential attachment in evolving networks", in: *EPL (Europhysics Letters)* 61.*4* (2003), p. 567.

[151] James Holland Jones and Mark S Handcock, "An assessment of preferential attachment as a mechanism for human sexual network formation", in: *Proceedings of the Royal Society of London B: Biological Sciences* 270.*1520* (2003), pp. 1123–1128.

[152] Uwe Deichmann, *A review of spatial population database design and modeling*, National Center for Geographic Information and Analysis, 1996.

[153] Huw Roland Jones, *Population geography*, Guilford Press, 1990.

[154] Deborah Balk and Gregory Yetman, "The global distribution of population: evaluating the gains in resolution refinement", in: *New York: Center for International Earth Science Information Network (CIESIN), Columbia University* (2004).

[155] Waldo Tobler, Uwe Deichmann, Jon Gottsegen, and Kelly Maloy, "World population in a grid of spherical quadrilaterals", in: *International Journal of Population Geography* 3.*3* (1997), pp. 203–225.

[156] Jerome E Dobson, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley, "LandScan: a global population database for estimating populations at risk", in: *Photogrammetric engineering and remote sensing* 66.*7* (2000), pp. 849–857.

[157] DL Balk, U Deichmann, G Yetman, F Pozzi, SI Hay, and A Nelson, "Determining global population distribution: methods, applications and data", in: *Advances in parasitology* 62 (2006), pp. 119–156.

[158] Catherine Linard, Marius Gilbert, Robert W Snow, Abdisalan M Noor, and Andrew J Tatem, "Population distribution, settlement patterns and accessibility across Africa in 2010", in: *PloS one* 7.*2* (2012), e31743.

[159] Andrea E Gaughan, Forrest R Stevens, Catherine Linard, Peng Jia, and Andrew J Tatem, "High resolution population distribution maps for Southeast Asia in 2010 and 2015", in: *PloS one* 8.*2* (2013), e55882.

[160] Derek Azar, Ryan Engstrom, Jordan Graesser, and Joshua Comenetz, "Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data", in: *Remote Sensing of Environment* 130 (2013), pp. 219–232.

[161] Uwe Deichmann, Deborah Balk, and Greg Yetman, "Transforming population data for interdisciplinary usages: from census to grid", in: *Washington (DC): Center for International Earth Science Information Network* (2001).

[162] Jeremy Mennis, "Generating Surface Models of Population Using Dasymetric Mapping?", in: *The Professional Geographer* 55.*1* (2003), pp. 31–42.

[163] FR Stevens, AE Gaughan, C Linard, and AJ Tatem, "Disaggregating census data for population mapping using random forests with remotely-sensed and other ancillary data", in: *PLoS One* (2014).

[164] Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L Urban, "LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics", in: *GeoJournal* 69.*1-2* (2007), pp. 103–117.

[165] Derek Azar, Jordan Graesser, Ryan Engstrom, Joshua Comenetz, Robert M Leddy Jr, Nancy G Schechtman, and Theresa Andrews, "Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti", in: *International Journal of Remote Sensing* 31.*21* (2010), pp. 5635–5655.

[166] Declan Butler, "Reactors, residents and risk", in: *Nature* 474 (2011), p. 36.

[167] Pinki Mondal and Andrew J Tatem, "Uncertainties in measuring populations potentially impacted by sea level rise and coastal flooding", in: *PloS one* 7.*10* (2012), e48191.

[168] Stephanie Wegscheider, Joachim Post, K Zosseder, M Mück, G Strunz, T Riedlinger, A Muhari, and HZ Anwar, "Generating tsunami risk knowledge at community level as a base for planning and implementation of risk reduction strategies", in: *Natural Hazards and Earth System Science* 11.*2* (2011), pp. 249–258.

[169] Marta M Jankowska, David Lopez-Carr, Chris Funk, Gregory J Husak, and Zoë A Chafe, "Climate change and human health: Spatial modeling of water availability, malnutrition, and livelihoods in Mali, Africa", in: *Applied Geography* 33 (2012), pp. 4–15.

[170] Andrew J Tatem, Nicholas Campiz, Peter W Gething, Robert W Snow, and Catherine Linard, "The effects of spatial population dataset choice on estimates of population at risk of disease", in: *Population health metrics* 9.*1* (2011), p. 4.

[171] Deepa K Pindolia, Andres J Garcia, Zhuojie Huang, David L Smith, Victor A Alegana, Abdisalan M Noor, Robert W Snow, and Andrew J Tatem, "The demographics of human and malaria movement and migration patterns in East Africa", in: *Malar J* 12.*397* (2013), pp. 10–1186.

[172] Andrew J Tatem, Susana Adamo, Nita Bharti, Clara R Burgert, Marcia Castro, Audrey Dorelien, Gunter Fink, Catherine Linard, John Mendelsohn, Livia Montana, et al., "Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation", in: *Popul Health Metr* 10.*8* (2012).

[173] Andrew J Tatem, "Mapping the denominator: spatial demography in the measurement of progress", in: *International health* (2014), ihu057.

[174] Samuel Leung, David Martin, and Samantha Cockings, "Linking UK public geospatial data to build 24/7 space-time specific population surface models", in: (2010).

[175] International Telecommunication Union, "World Telecommunication Development Conference (WTDC-2014): Final Report", in: (2014).

[176] Duncan J Watts, "A twenty-first century science", in: *Nature* 445.*7127* (2007), pp. 489–489.

[177] Alessandro Vespignani, "Predicting the behavior of techno-social systems", in: *Science* 325.*5939* (2009), p. 425.

[178] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson, "Approaching the limit of predictability in human mobility", in: *Scientific reports* 3 (2013).

[179] Olle Järv, Rein Ahas, Erki Saluveer, Ben Derudder, and Frank Witlox, "Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records", in: *PloS one* 7.*11* (2012), e49171.

[180] Maxime Lenormand, Miguel Picornell, Oliva G Cantú-Ros, Antònia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, and José J Ramasco, "Cross-checking different sources of mobility information", in: *PloS one* 9.*8* (2014), e105184.

[181] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti", in: *PLoS medicine* 8.*8* (2011), e1001083.

[182] Andrew J Tatem, Youliang Qiu, David L Smith, Oliver Sabot, Abdullah S Ali, Bruno Moonen, et al., "The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents", in: *Malar J* 8 (2009), p. 287.

[183] Andrew J Tatem, Zhuojie Huang, Clothilde Narib, Udayan Kumar, Deepika Kandula, Deepa K Pindolia, David L Smith, Justin M Cohen, Bonita Graupe, Petrina Uusiku, et al., "Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning", in: *Malaria journal* 13.*1* (2014), p. 52.

[184] Pablo Mateos and Peter F Fisher, "Spatiotemporal accuracy in mobile phone location: Assessing the new cellular geography", in: *Dynamic & Mobile GIS: Investigating Change in Space and Time* (2006), pp. 189–212.

[185] Statistics Portugal Instituto Nacional de Estatistica, "Population data from Portugal", in: (2011), URL: `www.ine.pt`.

[186] Institut National de la Statistique et des Etudes Economiques, "Population data from France", in: (2007), URL: `www.insee.fr`.

[187] 3GPP, "TS 03.02 V7.1.0 network architecture", in: (2000), URL: `www.3gpp.org/ftp/Specs/html-info/0302.htm`.

[188] Atsuyuki Okabe, Barry Boots, and Kokichi Sugihara, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, New York, NY, USA: John Wiley & Sons, Inc., 1992, ISBN: 0-471-93430-5.

[189] Andrew J Tatem *et al*, "WorldPop Project", in: (2015), URL: `www.worldpop.org.uk`.

[190] Autorite de Regulation des Communications Electroniques et des Postes (ARCEP), in: (2014), URL: `http://www.arcep.fr/`.

[191] Volker Bahn and Brian J McGill, "Testing the predictive performance of distribution models", in: *Oikos* 122.*3* (2013), pp. 321–331.

[192] R Statistical Package, "R: A language and environment for statistical computing", in: *Vienna, Austria: R Foundation for Statistical Computing* (2009).

[193] Alexander Brenning, "Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest", in: *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, IEEE, 2012, pp. 5372–5375.

[194] Mark Graham and Taylor Shelton, "Geography and the future of big data, big data and the future of geography", in: *Dialogues in Human Geography* 3.*3* (2013), pp. 255–261.

[195] Michael F Goodchild, "The quality of big (geo) data", in: *Dialogues in Human Geography* 3.*3* (2013), pp. 280–284.

[196] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang, "Modelling disease outbreaks in realistic urban social networks", in: *Nature* 429.*6988* (2004), pp. 180–184.

[197] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani, "The role of the airline transportation network in the prediction and predictability of global epidemics", in: *Proceedings of the National Academy of Sciences of the United States of America* 103.*7* (2006), pp. 2015–2020, DOI: 10.1073/pnas.0510525103, eprint: http://www.pnas.org/content/103/7/2015.full.pdf+html, URL: http://www.pnas.org/content/103/7/2015.abstract.

[198] Juan Camilo Bohorquez, Sean Gourley, Alexander R. Dixon, Michael Spagat, and Neil F. Johnson, "Common ecology quantifies human insurgency", in: *Nature* 462.*7275* (Dec. 2009), pp. 911–914, URL: http://dx.doi.org/10.1038/nature08631.

[199] James P. Bagrow, Dashun Wang, and Albert-László Barabási, "Collective Response of Human Populations to Large-Scale Emergencies", in: *PLoS ONE* 6.*3* (Mar. 2011), e17680, DOI: 10.1371/journal.pone.0017680, URL: http://dx.doi.org/10.1371.

[200] USAID, "Demographic and Health Surveys", in: (2014), URL: http://dhsprogram.com/.

[201] Amy Wesolowski, Nathan Eagle, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility", in: *Journal of The Royal Society Interface* 10.*81* (2013), ISSN: 1742-5689, DOI: 10.1098/rsif.2012.0986.

[202] Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility", in: *Sci. Rep.* 3 (Mar. 2013), URL: http://dx.doi.org/10.1038/srep01376.

[203] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási, "Structure and tie strengths in mobile communication networks", in: *Proceedings of the National Academy of Sciences* 104.*18* (2007), pp. 7332–7336.

[204] Gueorgi Kossinets and Duncan J Watts, "Empirical analysis of an evolving social network", in: *Science* 311.*5757* (2006), pp. 88–90.

[205] Lada A Adamic and Natalie Glance, "The political blogosphere and the 2004 US election: divided they blog", in: *Proceedings of the 3rd international workshop on Link discovery*, ACM, 2005, pp. 36–43.

[206] Michelle Girvan and Mark EJ Newman, "Community structure in social and biological networks", in: *Proceedings of the National Academy of Sciences* 99.*12* (2002), pp. 7821–7826.

[207] Santo Fortunato, "Community detection in graphs", in: *Physics Reports* 486.*3* (2010), pp. 75–174.

[208] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng, "Why we twitter: An analysis of a microblogging community", in: *Advances in Web Mining and Web Usage Analysis*, Springer, 2009, pp. 118–138.

[209] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E Tarjan, "Clustering social networks", in: *Algorithms and Models for the Web-Graph*, Springer, 2007, pp. 56–67.

[210] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai, "Lethality and centrality in protein networks", in: *Nature* 411.*6833* (2001), pp. 41–42.

[211] Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, et al., "A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae", in: *Nature* 403.*6770* (2000), pp. 623–627.

[212] Victor Spirin and Leonid A Mirny, "Protein complexes and functional modules in molecular networks", in: *Proceedings of the National Academy of Sciences* 100.*21* (2003), pp. 12123–12128.

[213] Jingchun Chen and Bo Yuan, "Detecting functional modules in the yeast protein–protein interaction network", in: *Bioinformatics* 22.*18* (2006), pp. 2283–2290.

[214] Jennifer A Dunne, Richard J Williams, and Neo D Martinez, "Network structure and biodiversity loss in food webs: robustness increases with connectance", in: *Ecology letters* 5.*4* (2002), pp. 558–567.

[215] Ann E Krause, Kenneth A Frank, Doran M Mason, Robert E Ulanowicz, and William W Taylor, "Compartments revealed in food-web structure", in: *Nature* 426.*6964* (2003), pp. 282–285.

[216] Albert-László Barabási, Réka Albert, and Hawoong Jeong, "Diameter of the world wide web", in: *Nature* 401.*9* (1999), pp. 130–131.

[217] Yon Dourisboure, Filippo Geraci, and Marco Pellegrini, "Extraction and classification of dense implicit communities in the web graph", in: *ACM Transactions on the Web (TWEB)* 3.*2* (2009), p. 7.

[218] Daniel Fenn, "Network communities and the foreign exchange market", PhD thesis, University of Oxford, 2010.

[219] Balachander Krishnamurthy and Jia Wang, "On network-aware clustering of web clients", in: *ACM SIGCOMM Computer Communication Review* 30.*4* (2000), pp. 97–110.

[220] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási, "Uncovering disease-disease relationships through the incomplete interactome", in: *Science* 347.*6224* (2015), p. 1257601.

[221] Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Thomas Couronné, Zbigniew Smoreda, and Carlo Ratti, "Delineating geographical regions with networks of human interactions in an extensive set of countries", in: *PloS one* 8.*12* (2013), e81707.

[222] Stanislav Sobolevsky, Izabela Sitko, Remi Tachet Des Combes, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti, "Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain", in: *Big Data (BigData Congress), 2014 IEEE International Congress on*, IEEE, 2014, pp. 136–143.

[223] Ruth Lapidoth, *Autonomy: Flexible solutions to ethnic conflicts*, US Institute of Peace Press, 1997.

[224] Svante E Cornell, "Autonomy as a source of conflict: Caucasian conflicts in theoretical perspective", in: *World politics* 54.*02* (2002), pp. 245–276.

[225] Gérard-François Dumont, *Les régions et la régionalisation en France*, Ellipses Paris, 2004.

[226] John J Beggs, Wayne J Villemez, and Ruth Arnold, "Black population concentration and black-white inequality: Expanding the consideration of place and space effects", in: *Social Forces* 76.*1* (1997), pp. 65–91.

[227] Charlie Karlsson and Michael Olsson, "The identification of functional regions: theory, methods, and applications", in: *The annals of regional science* 40.*1* (2006), pp. 1–18.

[228] Mike Coombes, "From city-region concept to boundaries for governance: The English case", in: *Urban Studies* (2013), p. 0042098013493482.

[229] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks", in: *Journal of Statistical Mechanics: Theory and Experiment* 2008.*10* (2008), P10008.

[230] Santo Fortunato and Claudio Castellano, "Community structure in graphs", in: *Computational Complexity*, Springer, 2012, pp. 490–512.

[231] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani, *The elements of statistical learning*, vol. 2, 1, Springer, 2009.

[232] James MacQueen et al., "Some methods for classification and analysis of multivariate observations", in: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, 14, Oakland, CA, USA., 1967, pp. 281–297.

[233] James C Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Kluwer Academic Publishers, 1981.

[234] Jörg Reichardt and Stefan Bornholdt, "Detecting fuzzy community structures in complex networks with a Potts model", in: *Physical Review Letters* 93.*21* (2004), p. 218701.

[235] Haijun Zhou, "Distance, dissimilarity index, and network community structure", in: *Physical review e* 67.*6* (2003), p. 061901.

[236] S Boccaletti, M Ivanchenko, V Latora, A Pluchino, and A Rapisarda, "Detecting complex network modularity by dynamical clustering", in: *Physical Review E* 75.*4* (2007), p. 045102.

[237] Mark EJ Newman, "Fast algorithm for detecting community structure in networks", in: *Physical review E* 69.*6* (2004), p. 066133.

[238] Philipp Schuetz and Amedeo Caflisch, "Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement", in: *Physical Review E* 77.*4* (2008), p. 046112.

[239] Juan Mei, Sheng He, Guiyang Shi, Zhengxiang Wang, and Weijiang Li, "Revealing network communities through modularity maximization by a contraction–dilation method", in: *New Journal of Physics* 11.*4* (2009), p. 043025.

[240] Mark EJ Newman and Michelle Girvan, "Finding and evaluating community structure in networks", in: *Physical review E* 69.*2* (2004), p. 026113.

[241] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral, "Modularity from fluctuations in random graphs and complex networks", in: *Physical Review E* 70.*2* (2004), p. 025101.

[242] Santo Fortunato and Marc Barthélemy, "Resolution limit in community detection", in: *Proceedings of the National Academy of Sciences* 104.*1* (2007), pp. 36–41.

[243] Josep M Pujol, Vijay Erramilli, and Pablo Rodriguez, "Divide and conquer: Partitioning online social networks", in: *arXiv preprint arXiv:0905.4918* (2009).

[244] Derek Greene, Donal Doyle, and Padraig Cunningham, "Tracking the evolution of communities in dynamic social networks", in: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, IEEE, 2010, pp. 176–183.

[245] Benjamin H Good, Yves-Alexandre de Montjoye, and Aaron Clauset, "Performance of modularity maximization in practical contexts", in: *Physical Review E* 81.*4* (2010), p. 046106.

[246] Marc Barthélemy, "Crossover from scale-free to spatial networks", in: *EPL (Europhysics Letters)* 63.*6* (2003), p. 915.

[247] Ling Heng Wong, Philippa Pattison, and Garry Robins, "A spatial model for social networks", in: *Physica A: Statistical Mechanics and its Applications* 360.*1* (2006), pp. 99–120.

[248] R. Lambiotte, V.D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical dispersal of mobile communication networks", in: *Physica A: Statistical Mechanics and its Applications* 387.*21* (2008), pp. 5317–5325, ISSN: 0378-4371.

[249] Marc Barthélemy, "Spatial networks", in: *Physics Reports* 499.*1* (2011), pp. 1–101.

[250] **Pierre Deville**, "Detection and analyses of communities in mobile phone networks", in: *UCL Master's thesis* (2011).

[251] Thomas Aynaud, Eric Fleury, Jean-Loup Guillaume, and Qinna Wang, "Communities in evolving networks: Definitions, detection, and analysis techniques", in: *Dynamics On and Of Complex Networks, Volume 2*, Springer, 2013, pp. 159–200.

[252] Theano S Terkenli, "Human activity in landscape seasonality: the case of tourism in Crete", in: *Landscape Research* 30.*2* (2005), pp. 221–239.

[253] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", in: *Nature* 435.*7043* (2005), pp. 814–818.

[254] Jeffrey Baumes, Mark K Goldberg, Mukkai S Krishnamoorthy, Malik Magdon-Ismail, and Nathan Preston, "Finding communities by clustering a graph into overlapping subgraphs.", in: *IADIS AC* 5 (2005), pp. 97–104.

[255] Jeffrey Baumes, Mark Goldberg, and Malik Magdon-Ismail, "Efficient identification of overlapping communities", in: *Intelligence and Security Informatics*, Springer, 2005, pp. 27–36.

[256] Vincenzo Nicosia, Giuseppe Mangioni, Vincenza Carchiolo, and Michele Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities", in: *Journal of Statistical Mechanics: Theory and Experiment* 2009.*03* (2009), P03024.

[257] Andrea Lancichinetti, Santo Fortunato, and János Kertész, "Detecting the overlapping and hierarchical community structure in complex networks", in: *New Journal of Physics* 11.*3* (2009), p. 033015.

[258] Anny Bloch and Alain Ercker, "Une culture de frontières entre l'Alsace et le Palatinat: Etat cruel des lieux", in: *Revue des sciences sociales de la France de l'Est 23* (1996), pp. 222–233.

[259] Andrés Rodríguez-Pose, "Social conditions and economic performance: The bond between social structure and regional growth in western Europe", in: *International Journal of Urban and Regional Research* 22.*3* (1998), pp. 443–459.

[260] James HS Bossard, "Residential propinquity as a factor in marriage selection", in: *American Journal of Sociology* (1932), pp. 219–224.

[261] C. Song, Z. Qu, N. Blumm, and A.L. Barabasi, "Limits of predictability in human mobility", in: *Science* 327.*5968* (2010), p. 1018.

[262] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel, "Unique in the Crowd: The privacy bounds of human mobility", in: *Scientific reports* 3 (2013).

[263] Duncan J Watts, *Six degrees: The science of a connected age*, WW Norton, 2004.

[264] Albert-László Barabási, "Linked: The New Science of Networks", in: (2002).

[265] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D Blondel, "D4D-Senegal: The Second Mobile Phone Data for Development Challenge", in: *arXiv preprint arXiv:1407.4885* (2014).

[266] Vittoria Colizza, Romualdo Pastor-Satorras, and Alessandro Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks", in: *Nature Physics* 3.*4* (2007), pp. 276–282.

[267] Vittoria Colizza and Alessandro Vespignani, "Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations", in: *Journal of theoretical biology* 251.*3* (2008), pp. 450–467.

[268] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi, "Collective response of human populations to large-scale emergencies", in: *PloS one* 6.*3* (2011), e17680.

[269] Liang Gao, Chaoming Song, Ziyou Gao, Albert-László Barabási, James P Bagrow, and Dashun Wang, "Quantifying information flow during emergencies", in: *Scientific reports* 4 (2014).

[270] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo, "Socio-Spatial Properties of Online Location-Based Social Networks.", in: *ICWSM* 11 (2011), pp. 329–336.

[271] Jon M Kleinberg, "Navigation in a small world", in: *Nature* 406.*6798* (2000), pp. 845–845.

[272] Marian Boguna, Dmitri Krioukov, and Kimberly C Claffy, "Navigability of complex networks", in: *Nature Physics* 5.*1* (2008), pp. 74–80.

[273] Marián Boguná, Fragkiskos Papadopoulos, and Dmitri Krioukov, "Sustaining the internet with hyperbolic mapping", in: *Nature communications* 1 (2010), p. 62.

[274] Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman, "Search in power-law networks", in: *Physical review E* 64.*4* (2001), p. 046135.

[275] Lada Adamic and Eytan Adar, "How to search a social network", in: *Social networks* 27.*3* (2005), pp. 187–203.

[276] Dashun Wang, Zhen Wen, Hanghang Tong, Ching-Yung Lin, Chaoming Song, and Albert-László Barabási, "Information spreading in context", in: *Proceedings of the 20th international conference on World wide web*, ACM, 2011, pp. 735–744.

[277] Everett M Rogers, *Diffusion of innovations*, Simon and Schuster, 2010.

[278] Eunjoon Cho, Seth A Myers, and Jure Leskovec, "Friendship and mobility: user movement in location-based social networks", in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2011, pp. 1082–1090.

[279] Jameson L Toole, Carlos Herrera-Yaqüe, Christian M Schneider, and Marta C González, "Coupling human mobility and social ties", in: *Journal of The Royal Society Interface* 12.*105* (2015), p. 20141128.

[280] Jean-Paul Rodrigue, Claude Comtois, and Brian Slack, *The geography of transport systems*, Routledge, 2013.

[281] Ed Bullmore and Olaf Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems", in: *Nature Reviews Neuroscience* 10.*3* (2009), pp. 186–198.

[282] Sven Erlander and Neil F Stewart, *The gravity model in transportation analysis: theory and extensions*, vol. 3, Vsp, 1990.

[283] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani, "The architecture of complex weighted networks", in: *Proceedings of the National Academy of Sciences of the United States of America* 101.*11* (2004), pp. 3747–3752.

[284] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman, "Power-law distributions in empirical data", in: *SIAM review* 51.*4* (2009), pp. 661–703.

[285] Mark EJ Newman and Juyong Park, "Why social networks are different from other types of networks", in: *Physical Review E* 68.*3* (2003), p. 036122.

[286] Réka Albert, Hawoong Jeong, and Albert-László Barabási, "Internet: Diameter of the world-wide web", in: *Nature* 401.*6749* (1999), pp. 130–131.

[287] Dirk Brockmann, Vincent David, and Alejandro Morales Gallardo, "Human mobility and spatial disease dynamics", in: *Reviews of nonlinear dynamics and complexity* 2 (2009), pp. 1–24.

[288] Dirk Brockmann and Dirk Helbing, "The hidden geometry of complex, network-driven contagion phenomena", in: *Science* 342.*6164* (2013), pp. 1337–1342.

[289] Romualdo Pastor-Satorras and Alessandro Vespignani, "Epidemic Spreading in Scale-Free Networks", in: *Phys. Rev. Lett.* 86 (14 2001), pp. 3200–3203, DOI: 10.1103/PhysRevLett.86.3200, URL: http://link.aps.org/doi/10.1103/PhysRevLett.86.3200.

[290] Marián Boguná and Romualdo Pastor-Satorras, "Epidemic spreading in correlated complex networks", in: *Physical Review E* 66.*4* (2002), p. 047104.

[291] Orion Penner, Raj K Pan, Alexander M Petersen, Kimmo Kaski, and Santo Fortunato, "On the predictability of future impact in science", in: *Scientific reports* 3 (2013).

[292] Filippo Radicchi, Santo Fortunato, and Claudio Castellano, "Universality of citation distributions: Toward an objective measure of scientific impact", in: *Proceedings of the National Academy of Sciences* 105.*45* (2008), pp. 17268–17272.

[293] Alexander Michael Petersen, Santo Fortunato, Raj K. Pan, Kimmo Kaski, Orion Penner, Armando Rungi, Massimo Riccaboni, H. Eugene Stanley, and Fabio Pammolli, "Reputation and impact in academic careers", in: *Proceedings of the National Academy of Sciences* (2014), DOI: 10.1073/pnas.1323111111, eprint: http:

//www.pnas.org/content/early/2014/10/03/1323111111.full.pdf+html, URL: http://www.pnas.org/content/early/2014/10/03/1323111111.abstract.

[294] Paul A David, "The Historical Origins of'Open Science': an essay on patronage, reputation and common agency contracting in the scientific revolution", in: *Capitalism and Society* 3.*2* (2008).

[295] R. Van Noorden, "Science on the move", in: *Nature* 490 (2012), p. 18.

[296] David Liben-Nowell and Jon Kleinberg, "The link-prediction problem for social networks", in: *Journal of the American society for information science and technology* 58.*7* (2007), pp. 1019–1031.

[297] Jinseok Kim, Jana Diesner, Heejun Kim, Amirhossein Aleyasen, and Hwan-Min Kim, "Why name ambiguity resolution matters for scholarly big data research", in: *Big Data (Big Data), 2014 IEEE International Conference on*, IEEE, 2014, pp. 1–6.

[298] Jinseok Kim and Jana Diesner, "The effect of data pre-processing on understanding the evolution of collaboration networks", in: *Journal of Informetrics* 9.*1* (2015), pp. 226–236.

[299] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation", in: *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, ACM, 2010, p. 4.

[300] James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, "Topic detection and tracking pilot study final report", in: (1998).

[301] https://publish.aps.org/datasets.

[302] Neil R Smalheiser and Vetle I Torvik, "Author name disambiguation", in: *Annual review of information science and technology* 43.*1* (2009), pp. 1–43.

[303] "Cultural Diversity", in: *HM Land Registry* (2006).

[304] Alan Porter and Ismael Rafols, "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time", in: *Scientometrics* 81.*3* (2009), pp. 719–745.

[305] Aron Culotta, Pallika Kanani, Robert Hall, Michael Wick, and Andrew McCallum, "Author disambiguation using error-driven machine learning with a ranking loss function", in: *Sixth International Workshop on Information Integration on the Web (IIWeb-07), Vancouver, Canada*, 2007.

[306] Vetle I Torvik, Marc Weeber, Don R Swanson, and Neil R Smalheiser, "A probabilistic similarity metric for Medline records: A model for author name disambiguation", in: *Journal of the American Society for information science and technology* 56.*2* (2005), pp. 140–158.

[307] Hui Han, Hongyuan Zha, and C Lee Giles, "Name disambiguation in author citations using a k-way spectral clustering method", in: *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, IEEE, 2005, pp. 334–343.

[308] José Soler, "Separating the articles of authors with the same name", in: *Scientometrics* 72.*2* (2007), pp. 281–290.

[309] Anderson A Ferreira, Adriano Veloso, Marcos André Gonçalves, and Alberto HF Laender, "Effective self-training author name disambiguation in scholarly digital libraries", in: *Proceedings of the 10th annual joint conference on Digital libraries*, ACM, 2010, pp. 39–48.

[310] Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations", in: *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, IEEE, 2004, pp. 296–305.

[311] Indrajit Bhattacharya and Lise Getoor, "A Latent Dirichlet Model for Unsupervised Entity Resolution.", in: *SDM*, vol. 5, 7, SIAM, 2006, p. 59.

[312] Hui Han, Wei Xu, Hongyuan Zha, and C Lee Giles, "A hierarchical naive Bayes mixture model for name disambiguation in author citations", in: *Proceedings of the 2005 ACM symposium on Applied computing*, ACM, 2005, pp. 1065–1069.

[313] William Cohen, Pradeep Ravikumar, and Stephen Fienberg, "A comparison of string metrics for matching names and records", in: *Kdd workshop on data cleaning and object consolidation*, vol. 3, 2003, pp. 73–78.

[314] Michael Levin, Stefan Krawczyk, Steven Bethard, and Dan Jurafsky, "Citation-based bootstrapping for large-scale author disambiguation", in: *Journal of the American Society for Information Science and Technology* 63.*5* (2012), pp. 1030–1047.

[315] Travis Martin, Brian Ball, Brian Karrer, and M. E. J. Newman, "Coauthorship and citation patterns in the Physical Review", in: *Phys. Rev. E* 88 (1 2013), p. 012814, DOI: 10.1103/PhysRevE.88.012814, URL: http://link.aps.org/doi/10.1103/PhysRevE.88.012814.

[316] Mark EJ Newman, "Coauthorship networks and patterns of scientific collaboration", in: *Proceedings of the National Academy of Sciences* 101.*suppl 1* (2004), pp. 5200–5205.

[317] Jiann-Wien Hsu and Ding-Wei Huang, "Correlation between impact and collaboration", in: *Scientometrics* 86.*2* (2011), pp. 317–324.

[318] Joseph C Lin, "Chinese names containing a non-Chinese given name", in: *Cataloging & classification quarterly* 9.*1* (1988), pp. 69–81.

[319] Shuk-fong Lau and Vicky Wang, "Chinese personal names and titles: Problems in cataloging and retrieval", in: *Cataloging & classification quarterly* 13.*2* (1991), pp. 45–65.

[320] Theresa A Velden, Asif-ul Haque, and Carl Lagoze, "Resolving author name homonymy to improve resolution of structures in co-author networks", in: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, ACM, 2011, pp. 241–250.

[321] "Everything you ever wanted to know about Korean surnames", in: *Korean Culture and Information Service, Retrieved 2015-01-18* (2015).

[322] http://nces.ed.gov/.

[323] Elsevier, *All Journals within Physics*, http://www.elsevier.com/journals/subjects/physics.

[324] Springer, *Journals in Physics*, http://www.springer.com/physics/journals.

[325] Wikipedia, *List of physics journals*, http://en.wikipedia.org/wiki/List_of_physics_journals.

[326] PhysNet, *Physics related free-access Journals*, http://de.physnet.net/PhysNet/journals.html.

[327] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency", in: *Advances in neural information processing systems* 16.*16* (2004), pp. 321–328.

[328] K Anders Ericsson, *The Cambridge handbook of expertise and expert performance*, Cambridge University Press, 2006.

[329] Alexander Fleming, "On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae", in: *British journal of experimental pathology* 60.*1* (1979), p. 3.

[330] Marie Curie, *Traité de radioactivité*, vol. 1, Gauthier-Villars, 1910.

[331] Emmy Noether, "Invariante variationsprobleme", in: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, mathematisch-physikalische Klasse* 1918 (1918), pp. 235–257.

[332] Francis Crick and James Watson, "Molecular structure of nucleic acids", in: *Nature* 171.*4356* (1953), pp. 737–738.

[333] S. Redner, "Citation Statistics from 110 Years of Physical Review", in: *Physics Today* 58 (2005), p. 49.

[334] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with Google PageRank algorithm", in: *Journal of Informetrics* 1.*1* (2007), pp. 8–15.

[335] A. Mazloumian, Y.H. Eom, D. Helbing, S. Lozano, and S. Fortunato, "How citation boosts promote scientific paradigm shifts and Nobel prizes", in: *PloS one* 6.*5* (2011), e18975.

[336] X. Shi, J. Leskovec, and D.A. McFarland, "Citing for high impact", in: *Proceedings of the 10th annual joint conference on Digital libraries*, ACM, 2010, pp. 49–58.

[337] A.M. Petersen, H.E. Stanley, and S. Succi, "Statistical regularities in the rank-citation profile of scientists", in: *Scientific reports* 1 (2011).

[338] Alexander M. Petersen, W.S. Jung, J.S. Yang, and H.E. Stanley, "Quantitative and empirical demonstration of the Matthew effect in a study of career longevity", in: *Proceedings of the National Academy of Sciences* 108.*1* (2011), pp. 18–23.

[339] Alexander M. Petersen, F. Wang, and H.E. Stanley, "Methods for measuring the citations and productivity of scientists across time and discipline", in: *Physical Review E* 81.*3* (2010), p. 036114.

[340] F. Radicchi and C. Castellano, "Rescaling citations of publications in physics", in: *Physical Review E* 83.*4* (2011), p. 046116.

[341] S. Redner, "How popular is your paper? An empirical study of the citation distribution", in: *The European Physical Journal B* 4.*2* (1998), pp. 131–134.

[342] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel, *Time series analysis: Forecasting and control*, Wiley. com, 2013.

[343] Benjamin Jones and Bruce Weinberg, "Age dynamics in scientific creativity", in: *Proceedings of the National Academy of Sciences* 108.*47* (2011), pp. 18910–18914.

[344] Laudeline Auriol, *Labour market characteristics and international mobility of doctorate holders: results for seven countries*, tech. rep., OECD Publishing, 2007.

[345] Laudeline Auriol, *Careers of doctorate holders: employment and mobility patterns*, tech. rep., OECD Publishing, 2010.

[346] R Van Noorden, "Global mobility: Science on the move", in: *Nature* 490 (2012), pp. 326–329.

[347] Quirin Schiermeier, "Career choices: The mobility imperative", in: *Nature* 470.*7335* (2011), pp. 563–564.

[348] Griet Jans et al., *Study on mobility patterns and career paths of EU researches*, tech. rep., European Commission, 2010.

[349] Andrés Solimano, *The International Mobility of Talent: Types, Causes, and Development Impact: Types, Causes, and Development Impact*, Oxford University Press, 2008.

[350] Michael Szell, Roberta Sinatra, Giovanni Petri, Stefan Thurner, and Vito Latora, "Understanding mobility in a social petri dish", in: *Scientific reports* 2.*457* (2012), DOI: 10.1.1/jpb001.

[351] Benjamin F Jones, Stefan Wuchty, and Brian Uzzi, "Multi-university research teams: Shifting impact, geography, and stratification in science", in: *science* 322.*5905* (2008), pp. 1259–1262.

[352] Andrejs Rauhvargers, "{Global university rankings and their impact}", in: *Leadership for WorldClass Universities Challenges for Developing Countries June* (2011).

[353] Isidro F Aguillo, Judit Bar-Ilan, Mark Levene, and José Luis Ortega, "Comparing university rankings", in: *Scientometrics* 85.*1* (2010), pp. 243–256.

[354] Qian Zhang, Nicola Perra, Bruno Gonçalves, Fabio Ciulla, and Alessandro Vespignani, "Characterizing scientific production and consumption in Physics", in: *Scientific reports* 3 (2013).

[355] Katy Börner and Shashikant Penumarthy, "Spatio-temporal information production and consumption of major US research institutions", in: *Proceedings of ISSI Volume* 1 (2005).

[356] Amin Mazloumian, Dirk Helbing, Sergi Lozano, Robert P Light, and Katy Börner, "Global Multi-Level Analysis of the ?Scientific Food Web'", in: *Scientific reports* 3 (2013).

[357] Harvey Goldstein and David J Spiegelhalter, "League tables and their limitations: statistical issues in comparisons of institutional performance", in: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (1996), pp. 385–443.

[358] Charles Eccles, "The use of university rankings in the United Kingdom", in: *Higher Education in Europe* 27.*4* (2002), pp. 423–432.

[359] Jorge E Hirsch, "Does the h index have predictive power?", in: *Proceedings of the National Academy of Sciences* 104.*49* (2007), pp. 19193–19198.

[360] Sune Lehmann, Andrew D Jackson, and Benny E Lautrup, "Measures for measures", in: *Nature* 444.*7122* (2006), pp. 1003–1004.

[361] Sharon G Levin and Paula E Stephan, "Are the foreign born a source of strength for US science?", in: *Science* 285.*5431* (1999), pp. 1213–1214.

[362] Lynne G Zucker and Michael R Darby, *Star scientists, innovation and regional and national immigration*, tech. rep., National Bureau of Economic Research, 2007.

[363] Chiara Franzoni, Giuseppe Scellato, and Paula Stephan, "Foreign-born scientists: mobility patterns for 16 countries", in: *Nature biotechnology* 30.*12* (2012), pp. 1250–1253.

[364] Mark EJ Newman, "Power laws, Pareto distributions and Zipf's law", in: *Contemporary physics* 46.*5* (2005), pp. 323–351.

[365] Staša Milojević, "Power law distributions in information science: Making the case for logarithmic binning", in: *Journal of the American Society for Information Science and Technology* 61.*12* (2010), pp. 2417–2425.

[366] Michel Feltin-Palas, "Réforme territoriale : L'homme qui a découpé les régions", in: (2004), URL: `http://www.lexpress.fr/region/l-homme-qui-a-dessine-les-regions_490366.html`.

[367] `http://genealogy.math.ndsu.nodak.edu/`.

[368] `http://en.wikipedia.org/wiki/Academic_genealogy_of_theoretical_physicists`.

[369] Ian Parberry and David S Johnson, "The SIGACT theoretical computer science genealogy: Preliminary report", in: (2004).

[370] Georgia T Chao, Patm Walz, and Philip D Gardner, "Formal and informal mentorships: A comparison on mentoring functions and contrast with nonmentored counterparts", in: *Personnel psychology* 45.*3* (1992), pp. 619–636.

[371] Declan Butler, "When Google got flu wrong.", in: *Nature* 494.*7436* (2013), p. 155.

[372] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, "The parable of Google Flu: traps in big data analysis", in: *Science* 343.*14 March* (2014).

[373] Jerry L Hintze and Ray D Nelson, "Violin plots: a box plot-density trace synergism", in: *The American Statistician* 52.*2* (1998), pp. 181–184.

[374] Robert McGill, John W Tukey, and Wayne A Larsen, "Variations of box plots", in: *The American Statistician* 32.*1* (1978), pp. 12–16.

[375] Bernard W Silverman, *Density estimation for statistics and data analysis*, vol. 26, CRC press, 1986.

[376] Student, "The probable error of a mean", in: *Biometrika* (1908), pp. 1–25.

[377] G Rupert Jr et al., *Simultaneous statistical inference*, Springer Science & Business Media, 2012.

[378] John W Tukey, "Comparing individual means in the analysis of variance", in: *Biometrics* (1949), pp. 99–114.

[379] Ben James Winer, Donald R Brown, and Kenneth M Michels, *Statistical principles in experimental design*, vol. 2, McGraw-Hill New York, 1971.

Appendix A

Statistical tools

In this first appendix, we present some of the statistical tools used in this work. The aim of this appendix is to not only provide a description of these tools but also to highlight their advantages over other tools and thus justifying our choices. We will first present the violin plot which is a method for plotting numerical data and efficiently visualising statistical distributions. We then describe the analysis of variance and the Tukey method which are statistical tests that we use to detect distributions that are significantly different from each other. These statistical tools are mainly used in Chapter 3.

A.1 Violin plot

First introduced in 1998 by *J. L. Hintze* and *R. D. Nelson* [373], the violin plot is a method for plotting statistical distribution. More particularly, it is combination of a box plot [374] and a rotated kernel density plot (or smoothed histogram) [375], which offers substantial improvements over other visualisation methods.

The box plot shows four distinct features about a variable and its distribution: center, spread, asymmetry and outliers. As shown in Figure A.1B, the labels identify the principal lines and points which form the structure of the traditional box plots: the median, the first and third quantile as well as the upper and lower adjacent value. In this work, the upper adjacent value is the largest observation that is less than or equal to the third quartile plus $1.5 \times IQR$, where IQR is the *interquartile range*, i.e. the difference between the first and third quartile. Similarly, the lower adjacent value is the smallest observation that is greater than or equal to the first quartile minus $1.5 \times IQR$.

As depicted in Figure A.1A, the violin plot includes a slightly modified version of the box plot. First, the median is represented by a circle which facilitates the comparison between different distributions. Second, the interquantile range is represented as a vertical thick black line, while the range between the upper and lower adjacent value is illustrated by a thiner vertical line. Finally, outliers are not represented by individual points.

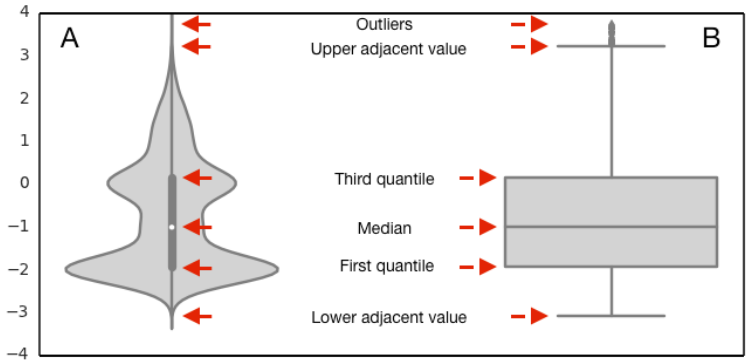


Figure A.1: Common components of (A) Violin plot and (B) Box plot, both representing the same sample distribution (n=50,000)

Besides integrating the main features of the box plot, the particularity of the violin plot is to graphically incorporate the distributional characteristics of the data, i.e. the probability density function. While a histogram could be useful for examining the distribution of a variable, the resulting graph can greatly differ depending on the number of intervals used. To overcome this problem, the violin plot uses a non-parametric density estimation called *kernel density* which estimates the probability density function of a variable based on the sample [375]. Less formally it can be seen as a way of averaging and smoothing a histogram. This estimation is essentially a sophisticated form of locally weighted averages of the distribution. By using a weight function, i.e. the kernel, an estimate of the density is created by placing a "bump" at each data point and then summing all bumps. Formally, given $(x_1, x_2, ..., x_n)$, an independent sample drawn from some distribution with an unknown density f, the kernel density estimator of f is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big) \tag{A.1}$$

where $K(.)$ is the kernel (a non-negative function which integrates to

one and has a mean equal to zero), x is the point where the density is estimated and $h > 0$ is the *bandwidth*, i.e. a smoothing parameter. In this work, we adopted the most commonly used kernel function, i.e. the normal density function, as well as the widely used *Silverman's rule of thumb* for the approximation of the bandwidth h [375], defined as

$$h = \left(\frac{4\sigma}{3n}\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-1/5} \tag{A.2}$$

and which is optimal if the underlying density being estimated is Gaussian but still robust for other distributions [375]. In Figure A.1A, the kernel density estimation is plotted symmetrically to the left and the right of the vertical box plot, allowing a quick and insightful comparison of several distributions.

With the addition of the probability density to the box plot , the violin plot provides a better description of the shape of the distribution; it highlights the peaks, valleys and bumps in the distribution. To demonstrate these advantages, we take as example random samples of 50,000 observations drawn from three distinct distributions: a normal, bimodal and uniform distribution. As depicted in Figure A.2, the box plots capture the fact that the three distributions share similar location and scale as measured by the median and the interquartile range, but rather fail to differentiate the shape between the three distributions, especially between the uniform and the bimodal.

On the other hand, the kernel density estimations on the violin plots accurately captures the different shapes (Figure A.3). For example, the violin plot for the bimodal clearly shows the twin peaks, while the corresponding box plot does not. As expected, it also well captures the fact that these three distributions share the same location and range.

Accurately capturing the shape of the distribution as the violin plot does is a fundamental aspect. Indeed, in an exploratory analysis, these plots enable the analyst to point to the next question which might investigate the reason behind particular shapes or trends detected in the data. In Chapter 3, these plots allow us to observe the two very different distributions of the parameters when choosing one cross-validation procedure over another (Fig. 3.6) or the different patterns of distributions of mobile phone user profiles over time (Fig. 3.14) which might be worth investigating in the future. All these observations could not have been made with a standard box plot method.

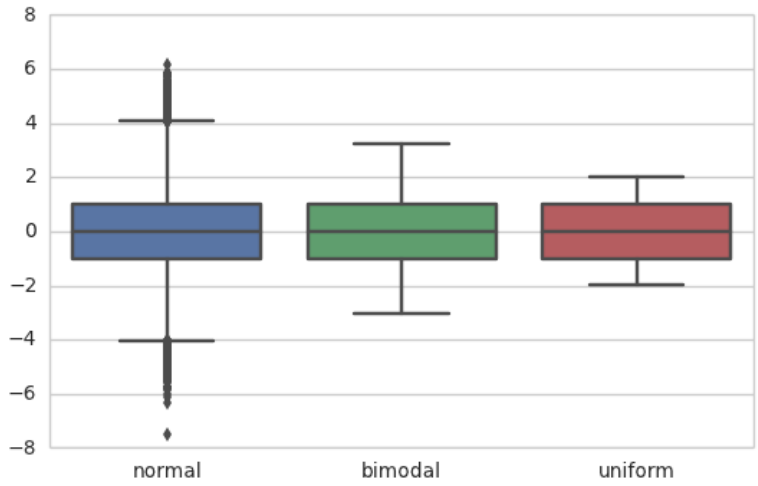


Figure A.2: Comparison of box plots for known distributions: normal, bimodal and uniform.

A.2 Analysis of variance

In Chapter 3 and more particularly in section 3.5, we are interested to compare and evaluate the precision (r) and accuracy ($RMSE$) of different types of mobile phone data; nightime, daytime, user-based and call-based data. To do so, we carried a cross-validation procedure where a sampling is repeated 1,000 times, providing precision and accuracy statistics for each data type independently (Fig. 3.11CD). In order to decide which data type is more accurate or precise than another, a statistical procedure is thus required to compare the four different distributions depicted in Figure 3.11CD and check wether one is significantly different than the others for both the precision and the accuracy.

A standard approach to assess wether two distributions are significantly different is the *2-sample t-test* [376]. In the case of two independent samples of equal size and similar variance, the t statistics can be calculated as follow

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \frac{1}{n}} \tag{A.3}$$

where X_1 and X_2 are the means of the two distributions, n the number of observations. Here, $s_{X_1X_2}$ is the grand standard deviation and is

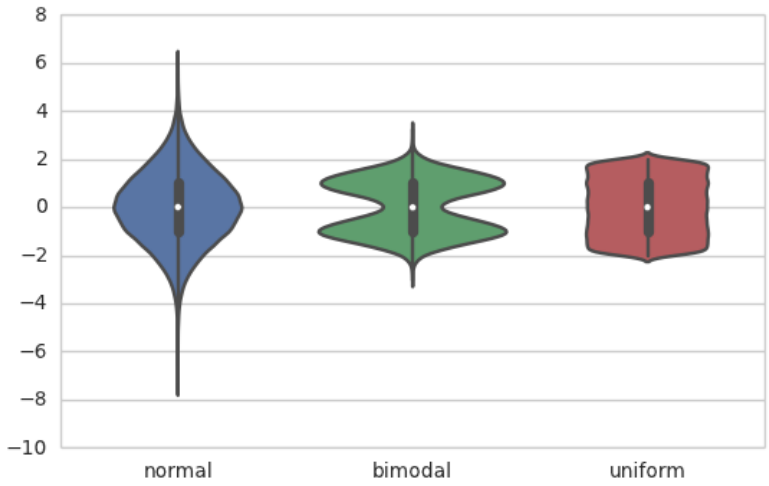


Figure A.3: Comparison of violin plots for known distributions: normal, bimodal and uniform.

defined as

$$s_{X_1X_2} = \sqrt{s^2_{X_1} + s^2_{X_2}} \tag{A.4}$$

where $s^2_{X_1}$ and $s^2_{X_2}$ are unbiased estimators of the variance of both distributions. For significance testing, the t statistic as well as the degree of freedom (n-2) are used to obtain the p-value to conclude wether the two distributions are the same (null hypothesis) or not.

In our case however, we do not have two distributions (or groups) but four (Fig. 3.11CD). A naive solution could be to perform pairwise t-tests between all groups. Unfortunately, multiple t-tests are not the answer because as the number of groups grows, the number of needed pairwise comparisons grows quickly, which induces statistical issues. Indeed, in the case of four groups, six pairwise comparisons need to be performed. Remember what a 0.05 significance (p-value=0.05) level means: you are willing to accept a 5% chance of a Type I error, rejecting the null hypotheses when it's actually true. But if you test six $p = 0.05$ hypotheses on the same set of data, you are more likely to commit a Type I error [377].

To resolve this issue, a common approach is the *analysis of variance* or ANOVA. This analysis provides a statistical test of whether or not the means of several groups are equal, and generalises the t-test to more than two groups. In this test, the null hypotheses is thus that all dis-

tributions are simply random samples of the same population. In our case, the null hypothesis is thus that the different data type have the same precision (Fig. 3.11C) or accuracy (Fig. 3.11D). Rejecting this hypothesis would imply that using different data types altered the accuracy or precision. Also, the ANOVA requires the samples to be random, independent as well as normally distributed and with similar variance. All these requirements are fulfilled by our experiment, even though the ANOVA is robust against these.

The question of the ANOVA test is thus whether the observed difference in means is too large to be the result of random selection. To do so, we look at the absolute difference of means between the groups, but we also consider the variability within each group. Intuitively, if the difference between groups is a lot bigger than the difference within groups, we conclude that the difference is not due to random and that there is a real effect. This is what the ANOVA does: comparing the variation between groups to the variation within groups. More formally, the following F-statistic is performed

$$F = MS_B/MS_W \tag{A.5}$$

corresponding thus to the ratio of the between-groups mean square MS_B over the within-groups mean square MS_W. This tells us how much more variability there is between treatment groups than within treatment groups. The larger that ratio, the more confident you feel in rejecting the null hypothesis, this hypothesis being that all means are equal and there is no significant difference. The value of the F-statistic as well as the number of groups and number of comparisons is then used to derive the p-value and thus to conclude wether the distributions are the same (the null hypothesis) or not. In Chapter 3, our experiments conclude that the input data has an effect on the accuracy ($F=989$, $p<0.01$, Fig. 3.11C) and precision ($F=368.9$, $p<0.01$, Fig. 3.11D), thus rejecting the null hypothesis.

Note that in the case where the null hypothesis is rejected, i.e. the distributions are not the same, we do not know which particular distribution is actually different from the others. To determine which one is significantly different, a *post-hoc analysis* in conjunction with the ANOVA is required.

A.3 Post-hoc analysis: Tukey test

If an analysis of variance (ANOVA) indicates that the distribution are different, the next step is often to determine *which* means are actually different. Such an analysis is called *post-hoc* analysis. In our case in Chapter 3, this analysis is particularly relevant as the analysis of variance indicated that the distributions were different for both the precision (Fig. 3.11C) and accuracy (Fig. 3.11D) but we were interested to detect *which* particular input data was significantly better than others.

There exists several different post-hoc analyses but the most common choice is the *Tukey* test or sometimes called the *Tukey's HSD* (honest significant difference) test [378]. This test is based on a similar formula than Eq. A.3 of the t-test, except that it corrects for the experiment-wise error rate explained before, i.e. the increasing chance of rejecting the null hypothesis while it is actually true. More precisely, the formula for the Tukey test is given by

$$q_t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{MS_W \cdot \frac{1}{n}}} \tag{A.6}$$

where X_1 and X_2 are the means of the two distributions, n the number of observations and MS_W the within-groups mean square, all obtained from the ANOVA. The q_t value can then be compared to the q value of a *studentized range* distribution [379] to determine wether the two distributions are the same (null hypothesis) or different.

Appendix B

Epidemic Spreading Simulations

B.1 Modeling

In Chapter 5, we identified a scaling law between social interactions and human mobility based on mobile phone data. In order to compare the accuracy and usefulness of these results, we simulated a Susceptible-Infected-Susceptible (SIS) process commonly used in modeling disease spreading [289, 290] by following the observed mobility fluxes T^M, the rescaled social fluxes $\tilde{T}^S$ but also the mobility fluxes T^M_{GM} approximated by the well-known gravity model [103, 105, 249, 282].

We consider the process where each location i (mobile phone tower) is characterized by a constant population size N_i equal to the number of distinct users present in the vicinity of the mobile tower over the period covered by the dataset D_1. The total population in our system is thus given by $\sum_{i=1}^{m} N_i$ and the system is equilibrated as the population is constant. In each location, users are classified according to their infectious state: they can be either infected (I) or susceptible to be infected (S). The standard generalization of this spatial SIS model is given by

$$S_i + I_i \xrightarrow{\mu} I_i \tag{B.1}$$

$$I_i \xrightarrow{\nu} S_i \tag{B.2}$$

$$S_i \xrightarrow{A_{i,j}} S_j \tag{B.3}$$

$$I_i \xrightarrow{A_{i,j}} I_j \tag{B.4}$$

where reaction (S5) indicates that susceptible users can become infected at a rate μ and reaction (S6) corresponds to infected users recovering from the disease at a rate ν. In addition to the standard SIS dynamics, susceptible as well as infected users can randomly move between one location i to another location j as described in reactions (S7) and (S8). The probability rate of these movements from location i to j is governed by the probability rate $A_{i,j}$ defined as

$$A_{i,j} = \frac{(T_{i,j}T_{j,i})^{-2}}{N_i}. \tag{B.5}$$

Since the system is equilibrated, the flux of users from i to j must balance that of j to i (detailed balance condition):

$$A_{i,j}N_i = A_{j,i}N_j \tag{B.6}$$

which is fulfilled by Eq. B.5.

In this case, the spatial SIS model can be defined as a set of m coupled ODEs for the infected people in each location [94, 266, 267, 287] :

$$\partial_t I_i = \mu \frac{I_i}{N_i}(N_i - I_i) - \nu I_i + \sum_{j \neq i}[A_{j,i}I_j - A_{i,j}I_i] \tag{B.7}$$

enabling us to compute the evolution of infected users in each location over time by solving these.